Available online at www.sciencedirect.com

ScienceDirect

Procedia Environmental Sciences 26 (2015) 38 – 44

Procedia Environmental Sciences

Spatial Statistics 2015: Emerging Patterns

Learning non-linear structures with Gaussian Markov random fields

Sara Fontanella[a]*, Lara Fontanella[b], Luigi Ippoliti[b], Pasquale Valentini[b]

[a] *Department of Mathematics and Statistics, The Open University, Walton Hall, Milton Keynes, MK7 6AA, UK*
[b] *Department of Economics, University of Chieti-Pescara, viale Pindaro, 42, 65127, Italy*

Abstract

Nowadays, one of the most changeling points in statistics is the analysis of high dimensional data. In such cases, it is commonly assumed that the dimensionality of the data is only artificially high: although each data point is described by thousands of features, it is assumed that it can be modeled as a function of only a few underlying parameters. Formally, it is assumed that the data points are samples from a low-dimensional manifold embedded in a high-dimensional space.
In this paper, we discuss a recently proposed method, known as Maximum Entropy Unfolding (MEU), for learning non-linear structures that characterize high dimensional data.
This method represents a new perspective on spectral dimensionality reduction and, joined with the theory of Gaussian Markov random fields, provides a unifying probabilistic approach to spectral dimensionality reduction techniques. Parameter estimation as well as approaches to learning the structure of the GMRF are discussed

Peer-review under responsibility of Spatial Statistics 2015: Emerging Patterns committee

*Keywords:*Nonlinear dimensionality reduction; Gaussian Markov random fields; Maximum Entropy Unfolding

* Corresponding author. Tel.: +44 190 865 2679.
E-mail address: Sara.Fontanella@open.ac.uk

Peer-review under responsibility of Spatial Statistics 2015: Emerging Patterns Committee
doi:10.1016/j.proenv.2015.05.020

1. Introduction

The aim of statistical methods for dimensionality reduction (DR) is to detect and discover low dimensional structures in high dimensional data. Many high-dimensional data in real-world applications can be modeled as data points lying close to a low-dimensional nonlinear manifold. The key observation is that, even if the dimension of the embedding spaces is very high, the intrinsic dimensionality of the data points could be rather limited. Traditional DR techniques, such as principal component analysis and multidimensional scaling, usually work well when the data points lie close to a linear (affine) subspace in the input space. They cannot, in general, discover nonlinear structures embedded in the set of data points.

Different methods have been proposed in literature for nonlinear dimensionality reduction[1].

In this paper we discuss a spectral nonlinear approach in the framework of Maximum Entropy Unfolding (MEU)[2]. It is shown that a low-dimensional representation of the data can be achieved through a spectral decomposition of the precision matrix of an Intrinsic Gaussian Markov random field (GMRF), which represents an important class of spatial models[3].

In general, the correlation structure of a GMRF model is hard to determine (except numerically), but the inverse correlations can be directly specified. On a lattice, this gives most, but not all, elements of the inverse dispersion matrix which is required for Gaussian maximum likelihood estimation of the proposed model. The Markov property of a GMRF makes it possible to utilize numerical methods for sparse matrices to construct fast algorithms for sampling and evaluation of the likelihood. Based on the theory of graphical models, we discuss both parameter estimation and possible approaches to retrieve the underlying structure (graph) of a GMRF.

The paper is organized as follows. Section 2 introduces the dimensionality reduction problem, while Section 3 provides details on MEU. In Section 4 we discuss parameter estimation based on Maximum Pseudolikelihood which, resulting in a least square estimator, favours the use of algorithms, such as Elastic Net and Lasso, to learn the graph structure of the Markov random field. Finally, Section 5 concludes the paper by discussing a Bayesian approach to parameter estimation.

2. The Manifold learning problem

Dimensionality reduction is the transformation of high-dimensional data into a meaningful representation of reduced dimensionality. Ideally, the reduced representation should have a dimensionality that corresponds to the *intrinsic dimensionality* of the data, defined as the minimum number of parameters needed to account for the observed properties of the data[4].

Specifically, let $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n]$ be a $(M \times n)$ data matrix of *input coordinates* consisting in n data-vectors, $\mathbf{x}_i \in \mathfrak{R}^M$, with $i = 1,2,\dots,n$. We further assume that the data matrix has *intrinsic dimensionality* m, with $m << M$.

In this context, the aim of dimensionality reduction methods is that of finding a $(m \times n)$ matrix of the coordinates in the reduced space, or *reconstructed embedding coordinate matrix*, $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n]$ with $\mathbf{y}_i \in \mathfrak{R}^m$, while retaining the geometry of the data as much as possible. The assumption that the observed data have intrinsic dimensionality m, implies that the data points lie on (or near) a manifold $\mathcal{M}$, with dimensionality m, that is embedded in the M-dimensional *input space* $\mathcal{X}$. To retrieve a faithful low dimensional representation of the data, which accounts for their intrinsic geometry, it is, then, important to determine the manifold metric.

Formally, the dimensionality reduction problem can be stated as follows: find a mapping $\Psi : \mathcal{M} \to \mathcal{Y} \subset \mathfrak{R}^m$, where $\mathcal{Y}$ is an affine space, called *feature space* of dimension $m < M$, such that the reconstructed embedding coordinates $\mathbf{Y}=\Psi(\mathbf{X})$ represent the metric structure in $\mathcal{X}$ well. In other words, given points $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\} \in \mathfrak{R}^M$ that lie on a m-dimensional manifold $\mathcal{M}$ that can be described by a single coordinate chart $f : \mathcal{M} \to \mathfrak{R}^M$, find $\{\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n\} \in \mathfrak{R}^m$, where $\mathbf{y}_i = f(\mathbf{x}_i)$[5].

Solving this problem is referred to as *manifold learning*, since we try to learn the manifold structure starting from a sample of data points. Hence, assuming the map Ψ to be non-linear allows capturing the nonlinearity of the underlying structure.

3. Maximum entropy unfolding

In recent years, several spectral nonlinear methods have been developed to address the manifold learning problem, including Isomap[6], Locally Linear Embedding[7], Laplacian eigenmaps[8], Kernel Principal Component Analysis[9] and Maximum Variance Unfolding[10].

Specifically, the spectral approach to dimensionality reduction involves the eigendecomposition of a $n \times n$ similarity matrix, also know as the Gram matrix, whose principal eigenvectors are extracted in order to retrieve a lower dimensional representation of the high dimensional data.

These nonlinear dimensionality reduction techniques are closely related and can be unified in a classical multidimensional scaling (CMDS)[11] perspective, where the Gram matrix is obtained from the pairwise squared Euclidean distances between the data points.

Starting form the perspective of CMDS, Lawrence[2] shows that the main difference between the nonlinear spectral DR techniques is in the distance matrices they use. Moreover, by following a different approach for constructing the distance matrix, Lawrence[2] proposes a new method, called the *Maximum Entropy Unfolding,* whose main feature is that of providing a probabilistic model. Exploiting the maximum entropy formalism[12], MEU specifies the probability density by a free form maximization of the entropy subject to constraints imposed on the expectations of the squared distances between two neighboring data points sampled from the model. For any two samples, $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, it thus holds

$$E(d_{p(\mathbf{X})}(i,j)) = d_{\mathbf{X}}(i,j) \;\; \forall\, j \;\in N(i)$$

where $N(i)$ represents the set of neighboring points of $\mathbf{x}_i$, and $d_{\mathbf{X}}(i,j)$ is the squared Euclidean distance between $\mathbf{x}_i$ and $\mathbf{x}_j$.

In order to maximize the entropy in a continuous system, Lawrence[2] proposes to maximize the negative Kullback-Leibler divergence (KLD), or relative entropy, between a base density, $q(\mathbf{X})$, and the density of interest, $p(\mathbf{X})$

$$\mathrm{H}=-\int \mathrm{p}(\mathbf{X}) \log_2 \frac{\mathrm{p}(\mathbf{X})}{\mathrm{q}(\mathbf{X})} \tag{1}$$

where $q(\mathbf{X})$ is defined as a very broad spherical Gaussian density with covariance $\gamma^{-1}\mathbf{I}$, with γ typically assumed to be close to zero. By writing the set of constraints as $\sum_{i=1}^{n}\sum_{j\,\in N(i)} \omega_{i,j} d_{\mathbf{X}}\,(i,j)$ where $\omega_{i,j}$ are the Lagrange multipliers, it can be shown that the probability distribution $p(\mathbf{X})$ corresponds to a zero-mean *Gaussian Markov random field* (GMRF)

$$p(\mathbf{X}) = \prod_{h=1}^{M} \frac{|\gamma\mathbf{I}+\mathbf{L}|^{1/2}}{(2\pi)^{n/2}} \exp\left\{-\frac{1}{2}\mathbf{x}_h'(\gamma\mathbf{I}+\mathbf{L})\mathbf{x}_h\right\} \tag{2}$$

with precision matrix $(\gamma\mathbf{I}+\mathbf{L})$. Note that, in order to capture the underlying structure of the manifold, $\mathbf{L}$ is defined as the Laplacian matrix, with off diagonal elements given by $-\omega_{i,j}$ if $\mathbf{x}_j$ is neighbor of $\mathbf{x}_i$ and 0 otherwise, and its diagonal elements defined as $\mathbf{L}(i,i) = \sum_{j\,\in N(i)} \omega_{i,j}$. Hence, $\mathbf{L}$ is symmetric and constrained to have a null space in the constant vector, $\mathbf{L1} = \mathbf{0}$, where $\mathbf{1}$ is the n-dimensional vector of ones.

Equation (2) enlightens that MEU implies independence across data features, which is due to the imposed constraints, and that the parameters, $\omega_{i,j}$, can be estimated through maximum likelihood. In fact, the gradient of each Lagrange multiplier is given by[2]

$$\frac{d\log p(\mathbf{X})}{d\omega_{i,j}} = \frac{1}{2}E_{p(\mathbf{X})}d(i,j) - \frac{1}{2}d_{\mathbf{X}}(i,j)$$

which can be evaluated by computing the expectation of the squared distance as

$$\mathrm{E}_{\mathrm{p}(\mathbf{X})}\big(\mathrm{d}(i,j)\big) = \mathrm{E}_{\mathrm{p}(\mathbf{X})}\big(\mathbf{x}_i'\mathbf{x}_i\big) + \mathrm{E}_{\mathrm{p}(\mathbf{X})}\big(\mathbf{x}_j'\mathbf{x}_j\big) - 2\mathrm{E}_{\mathrm{p}(\mathbf{X})}\big(\mathbf{x}_i\mathbf{x}_j\big) = \frac{\mathrm{M}}{2}[\mathrm{k}(i,i) + \mathrm{k}(j,j) - 2\mathrm{k}(i,j)] \tag{3}$$

where $k(i,j)$ is the (i,j)−th entry of the kernel (i.e. covariance matrix) $\mathbf{K} = (\gamma\mathbf{I} + \mathbf{L})^{-1}$. Equation (3) represents a scaled version of the standard transformation between distances and similarities. This relationship arises naturally in the probabilistic model since, in general, every GRF has an associated interpoint distance matrix and it is this matrix that is used in CMDS[2]. The parameterγ ensures that the precision matrix is positive definite. This implies that, knowing that the Laplacian has a null space in the constant vector, then $\mathbf{K1} = \gamma^{-1}$. This reflects an insensitivity of the covariance matrix to the data mean, and this in turn arises because that information is lost when we specify the expectation constraints only through interpoint distances. In practice, **K** is always centred before its eigenvectors are extracted, $\mathbf{S} = \mathbf{HKH}$, where the centering matrix is $\mathbf{H} = \mathbf{I} - \frac{1}{n}\mathbf{11}'$, with **I** a $(n \times n)$ identity matrix. Once the maximum likelihood n solution is recovered, the data can be visualized by looking at the eigenvectors of the Gram matrix **S**.

4. Parameter estimation

Optimal parameter estimation techniques, such as maximum likelihood (ML) estimation, can be computationally expensive, especially if noisy data are available and the noise is non-Gaussian. Also, there are many practical applications when fast solutions are required, even at the expense of a non-optimal estimate.

Hence, in light of equation (2), which tells us that our model is a GMRF, and according to graphical models theory, we propose a different approach for parameters estimation based on the maximization of the pseudolikelihood (PL) subject to constraints. Maximum pseudolikelihood (MPL) estimation provides an alternative, quick, and often reasonably efficient method of parameter estimation. Specifically, we discuss constrained MPL estimation to help deal with the problem of estimating Gaussian intrinsic autoregressions for which the model parameters are constrained to be on the boundary of the valid region.

In accordance with earlier discussion, we assume henceforth that each data point has a limited number of neighbors and that the conditional distribution of $\mathbf{x}_i$ is fully specified in terms of a vector $\boldsymbol{\theta}_i$ consisting of a few unknown parameters. Given the data, we shall use $p(\mathbf{x}_i|\mathbf{X}_{-i})$ to denote the conditional probability of observing $\mathbf{x}_i$, given all other values. The primary objective is to obtain a reasonable estimate of $\boldsymbol{\theta}_i$. The most naive approach to the estimation of the unknown parameters in the terms $p(\mathbf{x}_i|\mathbf{X}_{-i})$ would be to take the vector $\widehat{\boldsymbol{\theta}}_i$which maximizes the quantity

$$\Phi(\widehat{\boldsymbol{\theta}}_i) = \sum_{i=1}^{n} \log p\left(\mathbf{x}_i \middle| \mathbf{x}_j \; j \in N(i)\right) \tag{4}$$

with respect to $\boldsymbol{\theta}_i$.

From a graphical model perspective, this implies the specification of a family of probability distributions defined in terms of directed or undirected graphs. The nodes in the graph are identified with random variables, and joint probability distributions are defined by taking products over functions defined on connected subsets of nodes[13].

Although maximum pseudolikelihood estimation is intended to have fairly widespread applicability, it is of special interest in the Gaussian case in which we assume that

$$\mathbf{x}_i|\mathbf{x}_j \; , j \in N(i) \sim \mathcal{N}\left(\sum_{j \in N(i)} \beta_{i,j}\, \mathbf{x}_j \; , \tau_i^{-2}\mathbf{I}\right) \tag{5}$$

where τ_i^{-2} is the conditional variance.

According to Rue and Held[14], in the case of GRFs, the connected subsets of nodes are specified through the elements of the precision matrix, **L**. The pairwise conditional independence properties of $\mathbf{x}_i$ is contained solely in the covariance matrix and for detecting conditional independences, one must investigate **L**: each vertex represents a data point, and an undirected edge connects two data points, $\mathbf{x}_i$ and $\mathbf{x}_j$, if the corresponding element of the precision matrix, $\omega_{i,j}$, is non-zero. In order to analyse the nonzero pattern of the precision matrix $\mathbf{L} > 0$, and hence, the connectivity of the underlying graph we follow an approach based on the Cholesky decomposition of **L**:

$$\mathbf{L} = \mathbf{M}\mathbf{M}'$$

where $\mathbf{M}$ is a lower triangular matrix and represent a weighted adjacency matrix from a directed acyclic graph - DAG. This implies that when constructing the neighborhood, the triangular form for this matrix can be achieved by first imposing an ordering on the data points. Then, when seeking the nearest k neighbors for i, we only consider a candidate data point j if $j > i$. Applying recursively this procedure one vertex at time, we obtain that the i-th row of $\mathbf{M}$ provides an alternative parametrization of the conditional distribution of $\mathbf{x}_i|\mathbf{x}_j,\ j \in N(i)$. These conditional statements are described by a linear recursive system where the zero pattern for the regression coefficients is the same as in the concentration matrix. Hence, in light of these results and from equations (4) and (5), it can be shown that the estimated $\widehat{\boldsymbol{\beta}}$'s are the result of the solution of a system of recursive regressions

$$\mathbf{x}_i = (\mathbf{A}_i\mathbf{X}')'\boldsymbol{\beta}_i + \boldsymbol{\varepsilon}_i, \qquad i = 1,2,\dots,n \tag{6}$$

where $\mathbf{A}_i$ is a $(k \times n)$ indicator matrix which selects the k neighbors of $\mathbf{x}_i$. The $\widehat{\boldsymbol{\beta}}$'s are strictly related to the coefficients $\boldsymbol{\omega}$'s of the precision matrix $\mathbf{L}$ through

$$\beta_{i,j} = -\frac{\omega_{i,j}}{\omega_{i,i}}$$

MPL thus reduces to the ordinary method of least square which, for a design matrix $\widetilde{\mathbf{X}} = \mathbf{A}_i\mathbf{X}'$, gives

$$\widehat{\boldsymbol{\beta}}_i = \left(\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}}\right)^{-1}\widetilde{\mathbf{X}}'\mathbf{x}_i \tag{7}$$

and

$$\hat{\tau}^{-2} = \frac{1}{\mathrm{M}}\left[\left(\mathbf{x}_i - \widetilde{\mathbf{X}}\widehat{\boldsymbol{\beta}}_i\right)'\left(\mathbf{x}_i - \widetilde{\mathbf{X}}\widehat{\boldsymbol{\beta}}_i\right)\right] \tag{8}$$

is the estimate of the conditional variance.

4.1 Learning the graph structure

In the MEU framework, an important part of the model specification is the choice of the structure of the GMRF or, equivalently, of the precision matrix $\mathbf{L}$, which has the form of the Laplacian matrix.

As seen in the previous section, we parameterize the GMRF according to the Cholesky decomposition of $\mathbf{L}$, and, in order to constrain the Laplacian matrix to be positive semidefinite we may have to guarantee that assume that $\mathbf{M1} = \mathbf{0}$, which satisfy $\mathbf{L1} = \mathbf{0}$. If we force $\mathbf{M}(i,j) = 0$ if $j \notin N(i)$ and set the diagonal elements $\mathbf{M}(i,i) = -\sum_{\mathrm{j}\in\mathrm{N(i)}}\mathbf{M}(i,j)$, we will have a Laplacian matrix which is positive semidefinite without need of any further constraint on $\mathbf{M}$. One possibility to force a sparse structure in $\mathbf{M}$ is to work with a $k-$nearest neighbor rule, which assigns the same number of (nearest) neighbors to each data point.

The constraint $\mathbf{M1} = \mathbf{0}$ can be imposed by ensuring that sum of the off diagonal elements from each column is equal to one, i.e.$-\sum_{\mathrm{j}\in\mathrm{N(i)}}\beta_{ij} = 1, \quad i = 1,2,\dots,n$. Under general linear constraints we thus have that the negative pseudolikelihood can be reformulated as

$$\log \mathrm{p}(\mathbf{x}_i|\tau_i^{-2},\boldsymbol{\beta}_\mathrm{i}) \approx -\frac{\mathrm{M}}{2}\log\tau_i^{-2} + \frac{\tau_\mathrm{i}^{-2}}{2}\boldsymbol{\varepsilon}_i'\boldsymbol{\varepsilon}_i + \boldsymbol{\nu}(\mathbf{C}\boldsymbol{\beta}_i - \mathbf{c}) \tag{9}$$

where $\mathbf{C}$ is a $(r \times k)$ unit matrix which defines the linear system of the parameters which must be solved accordingly to the constraints in the r-vector $\mathbf{c}$, and $\boldsymbol{\nu}$ is a r-vector of Lagrangian multipliers. It is easy to show that the minimum of equation (9) is given by

$$\widetilde{\boldsymbol{\beta}}_i = \widehat{\boldsymbol{\beta}}_i - \mathbf{G}^{-1}\mathbf{C}'(\mathbf{C}'\mathbf{G}^{-1}\mathbf{C}')^{-1}(\mathbf{C}\boldsymbol{\beta}_i - \mathbf{c}) \tag{10}$$

where $\mathbf{G} = [\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}}]$. Equation (10) provides a constrained MPL estimator for the regression weights, which guarantees invariance to all the similarity transformations of the data.

The conditional variance, instead, is obtained as

$$\tilde{\tau}^{-2} = \frac{1}{M}\left[\left(\mathbf{x}_i - \widetilde{\mathbf{X}}\widetilde{\boldsymbol{\beta}}_i\right)'\left(\mathbf{x}_i - \widetilde{\mathbf{X}}\widetilde{\boldsymbol{\beta}}_i\right)\right] \quad (11)$$

4.1.1 The Elastic Net algorithm

An alternative procedure to learn the graph structure in $\mathbf{M}$ is that of using the Elastic Net (EN) algorithm. This is a regularization and variable selection method that adds the penalty on the L2-norm to the L1 penalty of the Lasso. The Elastic Net is particularly useful when the number of predictors (n) is much bigger than the number of observations (M), which is a case found in many examples. For any fixed non-negative λ_1 and λ_2 the elastic net criterion is defined as

$$H(\lambda_1, \lambda_2, \boldsymbol{\beta}_i) = |\mathbf{x}_i - \mathbf{X}_{-i}\boldsymbol{\beta}_i|^2 + \lambda_2|\boldsymbol{\beta}_i|^2 + \lambda_1|\boldsymbol{\beta}_i|_1 \quad (12)$$

where $\mathbf{X}_{-i}$ is the data matrix without the $i-$th input coordinate, $|\boldsymbol{\beta}_i|^2 = \sum_{j=1}^{n}\beta_{i,j}^2$ and $|\boldsymbol{\beta}_i|_1 = \sum_{j=1}^{n}|\beta_{i,j}|$. We refer to Hastie et al.[15] for know results on EN and its practical implementation.

5. Conclusion

In this paper, we considered MEU approach[2] that represents a novel approach to dimensionality reduction based on Gaussian random fields. The main advantage of MEU is that it provides a probabilistic model and a unifying perspective to spectral DR methods.

Given that we are required to estimate the parameters of a GMRF, we showed that, instead of applying the ML, one may use the PL approximation for obtaining these estimates. This is due to the conditional independence properties of GRMFs. We showed that the optimization of the PL turns out to be equivalent to solve a system of recursive equations, and hence we derived the maximum PL estimator through the least square method. Hence, PL allows for explicit estimation of the parameters through the regression formalism.

In order to guarantee the constraint concerning the null space of the Laplacian matrix, we derived also a constrained maximum PL estimator.

Moreover, in order to avoid user-defined parameters for the graph building, typical of rules like k-nn algorithm, we saw that combining PL with Elastic Net algorithm allows us to perform DR while learning the neighborhood relations directly from the data. Generally, by a good tuning of λ_1 and λ_2, EN can increase both sparsity and accuracy of the final solution.

Developing a Bayesian approach to achieve sparsity will be of interest for future works. One promising approach would be that of considering Bayesian spike and slab priors, in which the sparsity is induced by placing a mixture prior on the regression coefficient. Specifically, we will exploit the results obtained by George and McCulloch[16], who proposed a stochastic search variable selection (SSVS) algorithm for normal linear regression using Gibbs sampling to search for high posterior probability models. Their approach relies on a mixture of a low and high variance normal prior centered at zero for each of the regression coefficients, with the low variance component corresponding to a predictor being effectively excluded due to the coefficient being close to zero. For each independent regression in Equation (6), we suspect only a subset of the elements of $\boldsymbol{\beta}_i$ are non-zero. We assume that $\beta_{i,j}$ arises from one of two normal mixture components, depending on a latent variable $\gamma_{i,j}$:

$$\beta_{i,j}|\gamma_{i,j} \sim \gamma_{i,j}\mathcal{N}\left(\beta_{i,j}|0, \sigma^2\right) + (1-\gamma_{i,j})\mathcal{N}\left(\beta_{i,j}|0, c^2\sigma^2\right)$$

where σ is positive but small s.t. $\beta_{i,j}$ is close to zero when $\gamma_{i,j} = 0$; c is large enough to allow reasonable deviations from zero when $\gamma_{i,j} = 1$. In addition, the prior probability that $\mathbf{x}_j \in N(i)$ is:

$$P\left(\gamma_{i,j} = 1\right) = 1 - P\left(\gamma_{i,j} = 0\right) = p_k.$$

To obtain the normal mixture prior for $\boldsymbol{\beta}_i$, George and McCulloch[16] define a multivariate Normal prior

$$\boldsymbol{\beta}_i|\boldsymbol{\gamma}_i \sim \mathcal{N}_{\mathrm{M}}(\mathbf{0}, \mathbf{D}_{\boldsymbol{\gamma}_i}\mathbf{R}\mathbf{D}_{\boldsymbol{\gamma}_i})$$

where $\boldsymbol{\gamma}_i = (\gamma_{i,1}, \gamma_{i,2}, \dots, \gamma_{i,M})$ and $\mathbf{D}_{\boldsymbol{\gamma}_i} = \mathrm{diag}(\mathrm{d}_{k,1}\sigma, \mathrm{d}_{k,2}\sigma, \dots, \mathrm{d}_{k,M}\sigma)$ with $\mathrm{d}_{k,1} = 1$ if $\gamma_{i,1} = 0$ and $\mathrm{d}_{k,1} = \mathrm{c}$ if $\gamma_{i,1} = 1$.

References

1. Lee, J.A and Verleysen, M. *Nonlinear Dimensionality Reduction*. Springer; 2007.
2. Lawrence, N. D. A unifying probabilistic perspective for spectral dimensionality reduction: insights and new models. *Journal of Machine Learning Research* 2012; **13**:1609-1638.
3. Cressie, N. *Statistics for Spatial Data*. Wiley, New York; 1991.
4. Van der Maaten, L.J.P, Postma, E.O and Van den Herik, H.J. Dimensionality reduction: A comparative review. *Technical Report TiCC TR 2009-005*, 2009.
5. Cayton, L. Algorithms for manifold learning. *Univ. of California at San Diego Tech. Rep* 2005.
6. Tenenbaum J. B, De Silva V and Langford J. C. A global geometric framework for nonlinear dimensionality reduction. *Science* 2000; **290**:2319-2323.
7. Roweis, S. and Saul, L. Nonlinear dimensionality reduction by locally linear embedding. *Science* 2000; **290**:2323-2326.
8. Belkin, M. and Niyogi, P. Laplacian Eigenmaps for Dimensionality Reduction and Data Representation. *Neural Computation* 2003; **15**:1373-1396.
9. Scholkopf, B, Smola, A, Muller, K.R. Kernel principal component analysis. *Lecture Notes in Computer Science* 1998; **1327**:583-588.
10. Weinberger K.Q, Saul L.K. Unsupervised learning of image manifolds by semidefinite programming. *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, CVPR 2004. 988-995.
11. Mardia, K.V, Kent J.T and Bibby J.M. *Multivariate Analysis*. Academic Press, London; 1979.
12. Jaynes, E.T. Information Theory and Statistical Mechanics, *Phys.Rev*.1957; **106**:620-630.
13. Jordan, I. M. Graphical models. *Statistical Science* 2004; **19**(1):140155,.
14. Rue, H. and Held, L. *Gaussian Markov Random Fields: Theory and Applications*, volume 104 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, London, 2005.
15. Hastie,T, Tibshirani, R. and Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed., Springer, New York, 2009.
16. George, E. I. and Mcculloch, R. E. Variable Selection Via Gibbs Sampling. *Journal of the American Statistical Association* 1993; **88**(423):881-889