Literature Review

Thermography uses techniques to study heat radiations in different structures or regions. One of the unique example is its use in detecting tumors in human brain as corroborated by Gorbach et al.[47][48] and Shevelev et al.[49]. It analyses the heat radiated from brain tissues to determine whether it is a benign or a malign tissue. This greatly helps medical professionals to locate the bad tissues in brain. However, the radiations obtained from the surface is highly dynamic and non-stationary with continuous cerebral blood flow changes along with high environmental interference[50]. Therefore, it is of utmost priority to analyze the multidimensional data in such a way that it captures spatial as well as temporal interactions.

Before starting with the discussion on multidimensional analysis, it is crucial to discuss the existing methods for formulating univariate analysis i.e. analyzing the data in one dimension. The Generalized linear models (GLM) is one such method which were conceived by Nelder and Wedderburn[51] as a way of unifying various models, such that each outcome $\mathbf{Y}$ of the dependent variables are generated from a particular probability distribution that includes the normal, binomial, poisson, bernoulli and gamma distributions, among others. The mean, μ, of the distribution depends on the independent variables, $\mathbf{X}$, through:

$$\mathrm{E}(\mathbf{Y}) = \boldsymbol{\mu} = g^{-1}(\mathbf{X}\boldsymbol{\beta}) \tag{1}$$

where $\mathbf{E(Y)}$ is the expected value of $\mathbf{Y}$; $\mathbf{X}\beta$ is the linear predictor, β is a linear combination of unknown parameters ; g is the link function which provides the relationship between the linear predictor and the mean of the distribution function. The unknown parameters, β, are typically estimated with maximum likelihood technique which uses iterative re-weighted least squares algorithm[52]. Therefore, the GLM consists of three vital elements, a probability distribution from the exponential family. Secondly, a linear predictor $\eta = \mathbf{X}\beta$ and link function g such that $E(Y) = \mu = g^{-1}(\eta)$.

It is to be noted that GLM is always parametric for all practical purposes. This can be seen from the fact that GLM is stating linear relationships between variables, by $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}$. If a model is parametric, regression estimates the parameters from the data and if a model is linear in the parameters, estimation is based on methods from linear algebra that minimize the norm of a residual vector. Hence, the parametric regression model can be described using a finite number of parameters. These parameters are usually collected together to form a single n-dimensional parameter vector $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_n)$. Weisberg[1], or Draper and Smith[2] provided a detailed analysis of parametric regression method. However, a common shortcoming with this approach is that it is unable to catch random effects in the data which may portray non-linear relationships[3].

Hence, generalized linear model (GLM) is extended by non-parametric components which leads to semi-parametric regression or partially linear models[3]. These models combines the deterministic components with non-parametric components such as cubic B-Splines[4]. B-Spline function is a combination of flexible bands that passes through a number of points that are called control points or knots which creates smooth curves. These functions enable the creation and management of complex shapes and surfaces using an ideal number of knots. Generally speaking, cubic B-Splines which are third order polynomials normally is enough to define the curve and possesses high level of smoothness. However, choosing an optimal number and position of knots is a tricky task. Too many knots leads to overfitting of the data while too few knots leads to underfitting[5]. Friedman and Silverman[53] showed that increased flexibility provides smoothing procedure with an increased ability to fit the data. However, they discussed that this may be not good depending upon the extent to which the training sample is representative of the population of future observations to be predicted. Fitting the training data very closely results in degraded estimates with future performance. This phenomenon is hence called overfitting. Friedman and Silverman[53] quantified this phenomenon through bias-variance trade-off. They showed (future) expected squared error (ESE) as:

$$\mathrm{E}[f^\star(x) - f(x)]^2 = [f^\star(x) - Ef(x)]^2 + var f(x) \tag{2}$$

where f is a function to be estimated, $f^\star(x) = E(Y|X = x)$ is the conditional expectation for future observations with Y being the response variable and X being the observation variable. The expected values in above equation are overrepeated replicas of the training sample. $[f^\star(x) - Ef(x)]^2$ is the squared distance of the average (expected) curve estimate from the correct values defined as *bias squared* of the estimate. As the flexibility of the smoother increases, $[f^\star(x) - Ef(x)]^2$ decreases while $var f(x)$ increases. Therefore for each situation there is an optimal flexibility. Hence, to attain good performance while fitting the data, flexibility-continuity trade-off become vital[53].

P-Splines as proposed by Eilers and Marx[5] used a relatively large number of knots and a higher order finite difference penalty on coefficients of adjacent B-Splines. They introduced this idea of using difference penalty on coefficients after discussing the work of O'Sullivan[54] who used the integral of a squared higher derivative of the fitted curve as the penalty. P-Splines approach greatly reduces the dimensionality of the problem to n, the number of B-splines, instead of m, the number of observations, with smoothing splines[5]. Govindarajulu[55] compared performance of P-Splines with cubic B-Splines and natural splines for fitting non-linear exposure-response relationships with Cox Models. The P-splines had among the lowest Mean Square Error (MSE) when fitted to log or sine functions. The author concluded that applying penalized splines to exposure-response data provided the most consistent fit. The typical fit was good for all methods across all scenarios, but P-splines tended to exhibit the best behavior.

T.Hastie and R.Tibshirani[68] presented a penalized matrix decomposition (PMD), for computing a rank-K approximation for a matrix using L1-penalties, which yields a decomposition of $\boldsymbol{X}$ using sparse vectors. They show that when the PMD is applied using an L1-penalty on coefficients, a method for sparse principal components results. In fact, this yields an efficient algorithm for obtaining sparse principal components. The method is demonstrated on a publicly available gene expression data set. Andrew[69] studied two different regularization methods namely L1 and L2 for preventing overfitting. Focusing on logistic regression, he shows that using L1 regularization of the parameters, the sample complexity grows only logarithmically in the number of irrelevant features. This logarithmic rate matches the best known bounds for feature selection, and indicates that L1 regularized logistic regression can be effective even if there are exponentially many irrelevant features as there are training examples. They also give a lowerbound showing that any rotationally invariant algorithm including logistic regression with L2 regularization, has a worst case sample complexity that grows at least linearly in the number of irrelevant features.

T.Hastie and R.Tibshirani[56] developed a generalized additive model (GAM) approach to twist the properties of generalized linear models. A generalized additive model (GAM) is a generalized linear model with a linear predictor depends linearly on unknown smooth functions of some predictor variables. In general , the model has a structure something like:

$$g(\mathrm{E}(Y)) = \boldsymbol{X\beta} + f_1(x_1) + f_2(x_2) + \cdots + f_m(x_m). \tag{3}$$

where $\boldsymbol{X\beta}$ is the strictly parametric part of the model, f_i are the functions which may be specified parametrically, or semi-parametrically with smooth functions, Y is the univariate response variable whereas x_i are predictor variables. An exponential family distribution is specified for Y along with a link function g. The P-Splines combined with parametric component gives good approximation of the curve in a generalized additive model (GAM) approach. But this method cannot deal with numerical rank deficiency. Therefore, Wood[17] proposes GAM's with a ridge penalty and is based on the pivoted QR decomposition and the singular value decomposition. His method deals with rank deficiency even when numerical rank depends on smoothing parameters.

Let $\boldsymbol{X}$ be a $n \times q$ model matrix and $\boldsymbol{S_i}$ is the i_{th} (positive semidefinite) penalty matrix with unknown smoothing parameter θ_i. Wood[17] proposes to form the QR decomposition of $\boldsymbol{X}$,

$$\boldsymbol{X = QR} \tag{4}$$

where $\boldsymbol{Q}$ is made up of columns of an orthogonal matrix and $\boldsymbol{R}$ is upper triangular. Defining $\boldsymbol{S = H} + \sum_{i=1}^{m} \theta_i \boldsymbol{S_i}$ and $\boldsymbol{B}$ as any matrix square root of $\boldsymbol{S}$ such that $\boldsymbol{B^T B = S}$, a singular value decomposition can be formed:

$$\begin{bmatrix} \boldsymbol{R} \\ \boldsymbol{B} \end{bmatrix} = \boldsymbol{UDV_T} \tag{5}$$

$\boldsymbol{B}$ can be obtained efficiently by pivoted Choleski decomposition or by eigen-decomposition of the symmetric matrix $\boldsymbol{S}$. The columns of $\boldsymbol{U}$ are columns of an orthogonal matrix, $\boldsymbol{V}$ is an orthogonal matrix, and $\boldsymbol{D}$ is the diagonal matrix of singular values. Wood[17] says that examination of these singular values is the most reliable way to detect numerical rank deficiency of the fitting problem. In particular, at this stage any singular values that are too small should be removed along with the corresponding columns of $\boldsymbol{U}$ and $\boldsymbol{V}$. This deletion has the effect of recasting the problem into a reduced space in which the model parameters are identifiable. "Too small" is usually judged with reference to the largest singular value. In the work reported here, singular values less than the largest singular value multiplied by the square root of the machine precision were deleted. Therefore, this approach deals effectively with the difficult problem of rank deficiency that may occur only over part of the smoothing parameter space.

Various types of data comes as large grids of values. For example: data provided in the thermal imaging of the brain tissues, image data and multidimensional optical spectra often consists of data matrices of millions of points. Therefore, multivariate interpolation or spatial interpolation becomes important on functions of more than one variable[57]. Eilers and Marx[6] builds a two-dimensional coefficient surface that allows for interaction across the indexing plane of the regressor array. They presents a penalized signal regression using penalized B-Spline tensor products, where difference penalties are placed on rows and columns of the tensor product coefficients. The size of this model is not a problem but the intermediate step with flattened basis can lead to some huge issues. Imagine an image of 1000 by 1000 pixels and 1000 tensor products; the basis matrix would have 10^9 elements and consumes enormous amount of memory. Eilers, Currie and Durban[7] proposes a fast algorithm that takes advantage of the special structure of both the data as an array and the model matrix as a tensor product. This avoids computation of full basis matrix and computes the normal equations directly. This algorithm is therefore designed to handle huge basis functions.

Kernel smoothers[8] is another method which can be used to estimate a real valued function by its noisy observations, when no parametric model for this function is known. The estimated function is smooth, and the level of smoothness is set by a single parameter. However, kernel smoothers has one disadvantage that they have no parameters to characterize the coefficient surface. Bookstein[10] discusses thin plate splines as another approach to smooth two dimensional surface. The thin plate spline is the two-dimensional analog of the cubic spline in one dimension. It is the fundamental solution to the biharmonic equation. In principle, thin-plate splines (TPS) could be used as the model for multidimensional surfaces, but they have the problem of too many parameters to estimate with very large system of equations as shown by Wood[11]. Consider a non-parametric model as:

$$y_i = f(x_i) + \epsilon_i \tag{6}$$

The model above can be estimated by finding the function from an appropriate reproducing kernel Hilbert space which minimizes:

$$\parallel \boldsymbol{y} - \boldsymbol{f} \parallel^2 + \lambda \int f''(x)^2 dx \tag{7}$$

where $\boldsymbol{y}$ is a vector of $y_i's$, $\boldsymbol{f}$ is the corresponding vector of $f(x_i)$-values and $\parallel \Delta \parallel$ is the Euclidean norm and λ is a smoothing parameter. If we visualize this equation computationally, to fit a thin plate spline to n data points requires the estimation of n parameters and an additional smoothing parameter λ. Let d be the number of covariates. Except in the case $d = 1$ the above equation involves $O(n^3)$ operations, which is frequently prohibitive[11]. Additionally, thin plate spline fitting problems can have condition numbers in excess of 10^9, which has the potential to cause problems if a thin plate spline is embedded in a non-linear model, such as semi-parametric regressions for example.)

A way out could be to use thin-plate regression splines proposed by Wood[11]. The basis implied by solving the spline smoothing problem for a small representative data set of few hundred samples is found and this small basis is used to construct a model for the full data set of interest. However, both versions of TPS impose an isotropic penalty, which means that the amount of smoothing is the same for both dimensions. Normally, different types of spatial smoothers such as TPS discussed above uses euclidean distance between observations even though this distance may not be a measure of spatial proximity. Euclidean distance is extremely sensitive to the scales of the variables involved since all variables are measured in the same units of length. Secondly, the Euclidean distance is blind to correlated variables. The Mahalanobis distance is an alternative measure of the distance proposed by P.C.Mahalanobis[58]. The Mahalanobis distance of an observation $\vec{x} = (x_1, x_2, x_3, \ldots, x_N)^T$ from a set of observations with mean $\vec{\mu} = (\mu_1, \mu_2, \mu_3, \ldots, \mu_N)^T$ and covariance matrix S is defined as:

$$D_M(\vec{x}) = \sqrt{(\vec{x} - \vec{\mu})^T S^{-1} (\vec{x} - \vec{\mu})}. \tag{8}$$

The Mahalanobis distance takes into account the covariance among the variables in calculating distances. With this measure, the problems of scale and correlation inherent in the Euclidean distance are no longer an issue. If the covariance matrix is the identity matrix, the Mahalanobis distance reduces to the Euclidean distance. Using Mahalanobis distance in multidimensional data(example: spatial-temporal data) therefore seems a better choice as compared to euclidean distance.

The L-spline smoothing function proposed by Wahba[59] is a tool for estimating smooth univariate curves from data of the form $(x_1, y_1), \ldots\ldots., (x_n, y_n)$. The L-spline estimate of the curve represented by such a data set is the function $\hat{f}$ that minimizes the penalized sum-of-squares functional

$$\sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \int_a^b (Lf)^2 \tag{9}$$

subject to some appropriate boundary conditions. λ is a positive smoothing parameter and [a,b] is some interval which contains all of the x_i. The roughness penalty, L is defined by a linear differential operator of the form:

$$L = D^m + \omega_{m-1} D^{m-1} + ... + \omega_1 D + \omega_0 I \tag{10}$$

where D and I are the derivative and the identity operator respectively and the ω_i are arbitrary functions. The cubic smoothing spline is a special case of the L-spline, defined by $L = D^2$. Ramsay[13] derived a method named FELSPLINE (Finite element L Spline) in which he used a connection between smoothing with differential operators based penalties and partial differential equations to produce a smoother which solves partial differential equation problem defined only over finite area. However, a very strong boundary condition to ensure a unique solution to the L-spline smoothing is that the function $\hat{f}$ has zero normal derivative on the boundary of [a,b].

Wood[14] proposed an alternative by experimenting with the physical analogy of a soap film which can be represented as a basis penalty smoother, and has better boundary behavior. He imagined the domain boundary to be made of wire. This wire goes into a bucket of soapy water; a soap film with the same shape as the boundary is then formed. If the wire lies in the spatial plane, the height of the soap film at a given point is the value of the smooth at that point. Mathematically, the soap film consists of two sets of basis functions, one that is based entirely inside the domain and one that is induced by the (known or estimated) boundary values. One problem with soap film smoothing is that the basis setup is quite computationally expensive. A further problem is that no distinction exists between open boundaries (a boundary that is simply the limit of the region) and hard boundaries (real physical barriers).

Wang and Ranalli[15] replaces straight-line distances with geodesic distances in a smoother that is a sort of approximate thin plate spline (Geodesic Low rank Thin Plate Splines, GLTPS). This algorithm is cubic in the number of data, making the approach costly for large datasets. The principle difficulty in interpreting the results of their method is that it is unclear what their penalty term penalizes. To overcome these problems, Miller and Wood[16] uses a method of spline smoothing with respect to generalized distances proposed originally by Duchon[9]. They model a response y_i which is dependent on covariates via a linear predictor η_i:

$$\eta_i = \alpha_i + f(\boldsymbol{d_i}) \tag{11}$$

where α_i may depend linearly on further model coefficients (or may simply be zero). f is a smooth function, dependent on $\boldsymbol{d_i}$, a vector of generalized distances between the i^{th} observation. Finally they approximate the model as:

$$f(\boldsymbol{d_i}) = f_D(\boldsymbol{X}(\boldsymbol{d_i})) \tag{12}$$

where $\boldsymbol{x}(\boldsymbol{d})$ is the location of the point with distance vector $\boldsymbol{d}$ in the D dimensional Euclidean space. So the key idea here is that they smooth over a Euclidean space in which the Euclidean inter-observation distances are approximately equal to the original generalized distances. That is $\| \boldsymbol{x}(\boldsymbol{d_i}) - \boldsymbol{x}(\boldsymbol{d_j}) \| \approx d_{ij}$ when d_{ij} is the generalized distance between points i and j ($\|$ is the Euclidean norm).

Currie[12] introduced generalized linear array methods, or GLAM, in which data are arranged in an array structure or regular grid. GLAM model is based on generalized linear model (GLM) with the design matrix denoted as a Kronecker product. Suppose that the data $\mathbf{Y}$ is arranged in a $\mathbf{d}$-dimensional array with size $n_1 \times n_2 \times \ldots \times n_d$; thus,the corresponding data vector $\mathbf{y} = \mathbf{vec}(\mathbf{Y})$ has size $n_1 n_2 n_3 \cdots n_d$. Suppose also that the design matrix is of the form

$$\mathbf{X} = \mathbf{X}_d \otimes \mathbf{X}_{d-1} \otimes \ldots \otimes \mathbf{X}_1. \tag{13}$$

The standard analysis of a GLM with data vector $\mathbf{y}$ and design matrix $\mathbf{X}$ proceeds by repeated evaluation of the scoring algorithm

$$\mathbf{X}'\tilde{\mathbf{W}}_\delta \mathbf{X}\hat{\boldsymbol{\theta}} = \mathbf{X}'\tilde{\mathbf{W}}_\delta \tilde{\mathbf{z}}, \tag{14}$$

where $\tilde{\boldsymbol{\theta}}$ represents the approximate solution of $\boldsymbol{\theta}$, and $\hat{\boldsymbol{\theta}}$ is the improved value of it; $\mathbf{W}_\delta$ is the diagonal weight matrix, and $\mathbf{z} = \boldsymbol{\eta} + \mathbf{W}_\delta^{-1}(\mathbf{y} - \boldsymbol{\mu})$ is the working variable.

Computationally, GLAM provides array algorithms to calculate the linear predictor,

$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\theta} \tag{15}$$

and the weighted inner product $\mathbf{X}'\tilde{\mathbf{W}}_\delta \mathbf{X}$ without evaluation of the model matrix $\mathbf{X}$. thereby avoiding computational issues in storage and managing huge amount of data with high speed and efficient computations in model estimation.

It is straightforward to apply these ideas to the smoothing of multidimensional arrays. The author takes the marginal model matrix $\boldsymbol{X_i}$ to be $\boldsymbol{B_i}$, the smoother matrix for the i_{th} marginal variable. $\boldsymbol{B_i}$ is taken as regression matrix of B-splines. The overall smoother matrix $\boldsymbol{B}$ is then constructed as the Kronecker product of the $\boldsymbol{B_i}$. If $\boldsymbol{P}$ is the penalty matrix then equation (12) becomes

$$(\boldsymbol{B'\tilde{W}_\delta B} + \boldsymbol{P})\hat{\boldsymbol{\theta}} = \boldsymbol{B'\tilde{W}_\delta \tilde{z}} \tag{16}$$

To justify the claims, the author[12] uses three-dimensional American data on the number of deaths from respiratory disease. The data array $\boldsymbol{Y} = Y$[i,j,k] is indexed by age at death, i=1,.....,105, year of death, j =1,.......,40 (1959–1998), and month of death, k=1,......,12 and modeled with a GLAM. Thus $\boldsymbol{Y}$ has 50400 points arranged in a 105 × 40 × 12 array. The regression matrix $\boldsymbol{B}$ is defined via the marginal regression matrices of B-splines for age, $\boldsymbol{B_a}$, year, $\boldsymbol{B_y}$, and month, B_m. Hence, the full regression matrix looks like $\boldsymbol{B} = \boldsymbol{B_m} \otimes \boldsymbol{B_y} \otimes \boldsymbol{B_a}$ denoted as Θ. Assume λ_i are the smoothing parameters in each dimensions. The choice of initial values of smoothing parameters for the solution of above equation is extremely important. Therefore, they set $\lambda_a = \lambda_y = \lambda_m = 1$, say, and then iterate the above equation until convergence. This gives a second and much improved initial estimate of Θ which is used in the subsequent search for optimal values of $\lambda_a, \lambda_y, \lambda_m$. This alone cuts computer time by around 75% in this example. Also, the order of the storage of the variables made a difference to GLAM performance. The data has been stored with the largest variable (here age) varying fastest to smallest variable (here month) varying slowest. This ordering of the data implies that the regression matrix is $\boldsymbol{B} = \boldsymbol{B_m} \otimes \boldsymbol{B_y} \otimes \boldsymbol{B_a}$. This order is optimal in the sense that the number of multiplications that are performed to calculate the elements of $\boldsymbol{B'\tilde{W}_\delta B}$ is minimized. The algorithm takes about 25% longer to run with the variables in the reverse order.

A detailed discussion about existing methods on multi-dimensional smoothing has been done. However, analyzing spatial-temporal interactions specifically in the thermographic image dataset becomes key priority. Smoothing spatial-temporal model seems suitable to estimate simultaneously the spatial and temporal trends. Kammann and Wand[18] proposed geoadditive models with gaussian random fields where they imply that response variable is modelled as the sum of spatial and temporal effects as shown below:

$$f(space) + f(time). \tag{17}$$

The equation above doesn't count for space-time interaction effect. P-Spline ANOVA type interaction model for spatial-temporal smoothing proposed by Lee and Durban[19] allows spatial-temporal interactions. They used penalized splines in mixed model (semi-parametric regression) framework for smoothing spatial-temporal data with the model being as follows:

$$y = \gamma + f_s(x_1, x_2) + f_t(x_t) + f_{st}(x_1, x_2, x_t) + \epsilon, \epsilon \sim \mathcal{N}(0, \sigma^2). \tag{18}$$

where γ is linear predictor, ϵ is Gaussian error term with covariance being σ^2, $\boldsymbol{x_1}$ and $\boldsymbol{x_2}$ are spatial covariates, $\boldsymbol{x_t}$ is the temporal covariate, f_s is the two-dimensional spatial interaction, f_t is function to capture temporal trends and f_{st} is the spatial-temporal interaction. The spatial and spatial-temporal interactions uses two-dimensional and three-dimensional tensor products of P-Splines respectively. They applied constraints over the P-Spline regression coefficients, and therefore, the connection with the classical ANOVA decomposition is straightforward.

The literature of image analysis has recently seen an upsurge in the use of random field models to represent image data and to express prior, generic knowledge. A random field model (Besag[32][45]) assigns a color to each site, which either can represent intensity or range(depth) values or can denote a category label. A Markov random field (often abbreviated as MRF), is a set of random variables indexed by spatial positions. The work by (Abend et al. [43]) is probably the earliest work using the Markov assumption for pattern recognition. In MRF's, each pixel (often termed as a site) of an image act as a random variable having a Markov property described by an undirected graph. In other words, a random field is said to be Markov random field if it satisfies Markov properties[33]. A set of sites is assigned a set of labels. Let S index a discrete set of $n \times m$ sites in a rectangular lattice for a $2D$ image denoted by

$$S = \{(i, j) | 1 \leq i \leq n; 1 \leq j \leq m\} \tag{19}$$

Its elements correspond to the locations at which an image is sampled. The inter-relationship between sites is maintained by a so-called *neighborhood* system. A label is an event that may happen to a site. Let L be a set of labels. A label set may be categorized as being continuous or discrete. In terms of the regularity and the continuity, a vision labeling problem is classified into $LP1$: Regular sites with continuous labels and $LP2$: Regular sites with discrete labels. MRF's with discrete labels are known as Discrete Markov Random fields whereas MRF's with continuous labels are known as Gaussian Markov Random fields. The sites in S are related to one another via a neighborhood system. A neighborhood system for S is defined as

$$N = \{N_i | \forall_i \varepsilon S\} \tag{20}$$

where N_i is the set of sites neighboring i. In the first order neighborhood system, also called the 4-neighborhood system, every (interior) site has four neighbors. When the sites in a regular rectangular lattice correspond to the pixels of an $n \times m$ image in the $2D$ plane, an internal site (i,j) has four nearest neighbors as $N_{i,j} = \{(i-1, j), (i+1, j), (i, j-1), (i, j+1)\}$, a site at a boundary has three and a site at the corners has two. The pair $(S, N) \equiv G$ constitutes a graph in the usual sense; S contains the nodes and N determines the links between the nodes according to the neighboring relationship. A clique C for (S, N) is defined as a subset of sites in S. The single-site and horizontal and vertical pair-site cliques constitutes the first order neighborhood system. Let $F = \{F_1, ..., F_m\}$ be a family of random variables defined on the set S, in which each random variable F_i takes a value f_i in L. The family F is called a random field. We use the notation $F_i = f_i$ to denote the event that F_i takes the value f_i and the notation $(F_1 = f_1, ..., F_m = f_m)$ to denote the joint event. For a discrete label set L, the probability that random variable F_i takes the value f_i is denoted $P(F_i = f_i)$, abbreviated $P(f_i)$ and the joint probability is denoted $P(F = f) = P(F_1 = f_1, ..., F_m = f_m)$ and abbreviated $P(f)$.

F is said to be a Markov random field on S with respect to a neighborhood system N if and only if the following two conditions are satisfied:

$$P(f) > 0, \forall f \varepsilon F \qquad (positivity) \tag{21}$$

$$P(f_i | f_{S-i}) = P(f_i | f_{N_i}), \qquad (Markovianity) \tag{22}$$

Markovianity can be termed as the property of a random variable, if the conditional probability distribution of the random variable depends only upon its neighboring variables N_i instead of all the random variables present in the system $S - i$.

Gibbs Random Field (GRF) is another type of random field that needs to be discussed to model a random field. A set of random variables F is said to be a Gibbs random field (GRF) on S with respect to N if and only if its configurations obey a Gibbs distribution. A Gibbs distribution takes the following form

$$P(f) = Z^{-1} \times e^{-\frac{1}{T}U(f)} \tag{23}$$

where

$$Z = \sum_{f \epsilon F} e^{-\frac{1}{T}U(f)} \tag{24}$$

is a normalizing constant called the partition function, T is a constant called the temperature which shall be assumed to be 1 unless otherwise stated, and $U(f)$ is the energy function. The energy

$$U(f) = \sum_{c \epsilon C} V_c(f) \tag{25}$$

is a sum of clique potentials $V_c(f)$ over all possible cliques C. The value of $V_c(f)$ depends on the local configuration on the clique c. An important special case is when only cliques of size up to two i.e first order are considered. In this case, the energy can also be written as

$$U(f) = \sum_{i \epsilon S} V_1(f_i) + \sum_{i \epsilon S} \sum_{i' \epsilon N_i} V_2(f_i, f_{i'}) \tag{26}$$

An MRF is characterized by its local property (the Markovianity) whereas a GRF is characterized by its global property (the Gibbs distribution). The Hammersley-Cliord theorem (Hammersley and Cliord [44]) establishes the equivalence of these two types of properties. Besag[32] and Kindermann and Snell[33] justifies Hammersley-Cliord theorem that a unique GRF exists for every discrete MRF field and visa-versa as long as Gibbs random field is defined in terms of a neighborhood system. The reports by Cross and Jain [35], Geman and Geman [36], Cohen and Cooper [37], and studies by Derin and Elliott [34][38] all make use of the Gibbs distributions (GD) for characterizing MRF.

Kindermann and Snell[33] also shows that a discrete Gibbs random field (GRF) provides a global model for an image by specifying a probability mass function. It describes the global properties of an image in terms of the joint distribution of colors for all pixels. Derin and Elliot[34] presented a new approach to the use of Gibbs distributions. They proposed dynamic programming based segmentation algorithms for noisy and textured images, considering a statistical maximum a posteriori (MAP) criterion. Pappas[46] generalized K-means clustering algorithm to include spatial constraints using eight neighbor Gibbs random field applied to pictures of industrial objects, buildings, aerial photographs, optical characters, and faces. It shows that the algorithm performs better than the K-means algorithm by the use of

Gibbs random fields. Gurelli and Onural[47] talks about a Gibbs-Markov random field (GMRF) parameter estimation technique proposed by Derin and Elliott[34]. They refer this technique as the histogramming (H) method. First, the relation of the H method to the (conditional) maximum likelihood (ML) method is considered. Second, a bias-reduction based modification of the H method is proposed to improve its performance.

The ubiquity of Gaussian random variables in statistical applications has led to the use of continuous random fields, called Gaussian Markov Random Fields (GMRF) as image models. Chellappa[39] proposed the algorithm for sampling GMRFs. Pixel or sites are assigned continuous labels which have joint Gaussian distributions with means μ, standard deviations σ, and correlations controlled by $\boldsymbol{\theta}$ parameters. Therefore, GMRF is a pairwise interaction model with the color at each pixel permitted to take on any real value. Let $P(F_i = f_i)$ be the probability that a random variable F_i takes a real value f_i given by:

$$P(F_i = f_i) = \frac{(x_i - \mu)^2}{2\sigma^2} \tag{27}$$

So, the pairwise interaction model with the neighboring pixels or sites becomes

$$P(f_i|f_{N_i}) = -\boldsymbol{\theta}\frac{(x_i - \mu)(x_{i;N_i} - \mu_{i;N_i})}{\sigma^2} \tag{28}$$

For purposes of sampling a GMRF, $n \times m$ random variables are viewed representing pixel colors. The covariance matrix is $\sigma_2\boldsymbol{B^{-1}}$ where $\boldsymbol{B^{-1}}$ is a correlation matrix. The inverse of the correlation matrix is called precision matrix, $\boldsymbol{B}$ and is given by

$$\begin{bmatrix} B_{11} & B_{12} & B_{13} & \dots & B_{1m} \\ B_{21} & B_{22} & B_{23} & \dots & B_{2m} \\ \dots & \dots & \dots & \dots & \dots \\ B_{n1} & B_{n2} & B_{n3} & \dots & B_{nm} \end{bmatrix}$$

where $B_{ij} = 0$ whenever the sites are not neighbors. This produces a sparse precision matrix but the inverse of this matrix i.e correlation matrix becomes dense. A major difficulty with using GMRF models is the selection of parameters θ for which the correlation matrix B^{-1} is positive definite. Without this property, the inverse doesn't exist and the model has no standing. This is a serious difficulty in parameter estimation. With discrete GRFs and MRFs, almost any parameter values will generate mathematically valid models. However, only a portion of the potential parameter space leads to valid GMRF models. The region of validity is known explicitly for first order GMRFs but is not known for higher order GMRFs. The parameters of a GMRF can be estimated in several ways, such as by Besag's coding scheme (Besag[32]), by pseudo-likelihood maximization (Besag[40][41]), by minimizing the sum of square errors (Chellappa[39]) and by maximum likelihood estimation. Manjunath and Chellappa[42] used the fourth-order GMRF to model the conditional probability density of the image intensity array given its texture labels. The texture labels were assumed to obey a first- or second-order discrete Markov model with a single parameter β, which measures the amount of clustering between adjacent pixels.

Fahrmeir, Kneib and Lang[20][22][23][24][25][26][27][28] proposed extension of penalized spline semi-parametric space-time regression model using a Bayesian perspective. The non-linear effects of continuous co-variates and time trends are modeled using Bayesian versions of penalized splines, while correlated spatial effects has been solved using Gaussian Markov random field prior working in continuous schema. Inference has been performed using Empirical Bayes approach on the basis of generalized linear mixed model representation. This approach of inference has been termed as posterior mode estimation and resembles penalized likelihood estimation. The advantage of Bayesian approach is that all unknown functions and parameters can be in a unified framework by assigning appropriate prior probabilities. The empirical approach is based on generalized linear mixed model (GLMM) developed by Lin and Zhang[21] for longitudinal data analysis using smoothing splines, or in Kammann and Wand[18] for geoadditive models using stationary Gaussian random fields. The spatial-temporal model proposed by them is as follows:

$$\eta_{it} = f_1(x_{it1}) + ... + f_k(x_{itk}) + f_{time}(t) + f_{spat}(s_{it}) + u'_{it}\gamma \tag{29}$$

where η_{it} are the predictor values, $x_{it1}, ..., x_{itk}$ are covariate values, f_{time} are possibly non-linear time trend, f_{spat} is a spatially correlated random effect of the location s_{it} an observation pertains to. γ is the regression parameter and $u'\gamma$ is the linear predictor. The time trends and the continuous covariates above has been modelled using P-Splines and the spatial effect by Gaussian Markov random field which works in continuous schema.

In order to check the performance of Empirical Bayes (EB) approach using Gaussian Markov Random field compared to Full Bayes (FB) approach using Markov Chain Monte Carlo (MCMC) technique, the author[20] uses the longitudinal data which is collected in yearly visual forest damage inventories carried out in a forest district in the northern part of Bavaria from 1983 to 2001. The spatial-temporal model using EB approach shows a clear improvement as compared to FB approach, confirming that inclusion of the spatial information is substantial. Therefore, the author concludes that Empirical Bayes inference is a promising alternative to full Bayes inference even for fairly large data sets. Biller[29], Gamerman[30], Lang[31] also proposed Full Bayesian approach to inference using Monte Carlo technique however Monte Carlo simulations are computationally very expensive as compared to Empirical Bayes approach using Markov random field priors.

Till now we have seen that, a common method to smooth a spatially distributed process is to employ Markov Random field (MRF) prior. The prior utilizes neighboring values based on predefined neighborhood structure. In popular literature, the discrete Markov Random fields is subdivided into Ising prior which is having binary labels and Potts prior with a range of discrete values. The other MRF priors can be modeled using Gaussian MRFs (Besag[32]). However, there have been very few attempts in the literature to make a comparison of the properties, complexities as well as performances of these priors. Consequently, statisticians blindly chooses between them without any justification. Smith and Smith[60] made a valiant effort to compare Discrete MRFs based on Ising prior with Gaussian MRFs in the context of two different empirical examples. All examples feature a two-dimensional regular lattice and for both type of priors the neighborhood structure involves eight immediate neighboring pixels.

The first example is the 20^{th} of 32 progressively deeper slices from a confocal fluorescence microscopy acquisition on a stomach leaf cell as analyzed by Hurn[61]. A laser recorded fluorescence (range [0,255]) at each pixel in a 170×127 lattice. The posterior probability map and the cross section in Figures in [60] show how the Ising prior smooths over this gap, estimating high posterior probabilities of the pixels as compared to Gaussian MRFs. The tendency of the Ising prior to fill in voids and smooth over small gaps is also noticeable in the example. Overall, the Ising prior appears to promote stronger clustering than the Gaussian prior. The author[60] also measured the time complexity of both the priors. Precomputations for Gaussian prior took approximately four hours whereas it took approximately 3 minutes per grid to compute the Ising prior using the same conservative sampling periods[60].

The second example in [60] was drawn from the literature on spatial-temporal modeling of functional magnetic resonance imaging (fMRI) data in human brain imaging. The data had high levels of noise an spatial correlation and spatial smoothing was crucial to obtain quality results. So the fMRI was fitted using the two priors followed by the simulation study based on the fitted data. The objective in fMRI studies of the human brain was to identify areas where an increase in the blood oxygenation levels occurs in response to the presence of an external stimulus, as measured by an observed fMRI signal time series at each site (labeled a "voxel") on a large regular three-dimensional lattice. The local increase of blood oxygenation is associated with neuronal activity, which is known to occur in regions of the brain corresponding to clusters of spatially contiguous voxels. The dataset used here is voxels of the brain in one 72×86 lattice. The data for this study are derived from time series regressions, each with 63 observations of the noisy fMRI signal as the dependent variable, carried out at each voxel. The Discrete MRFs i.e Ising prior and Gaussian MRFs are then used to smooth the data. The author[60] founds that there are subtle differences on how the two different priors performs smoothing. The Ising prior has less "starring" (isolated voxels being classified as active) than the Gaussian prior, and features more clustering of active voxels than Gaussian prior. To investigate further, the author undertakes a simulation study on fitted regressions. The noise level of the regression is set equal to a multiple $c\epsilon 0.5, 1, 2$ of the estimated noise in each regression. At each voxel, 63 observations of the dependent variable are generated. Then, using each of the three different priors, classifications were obtained. This process was repeated to obtain 50 replicates for the simulation. The average of the metrics at each noise level over 50 replications for each template are shown in Table.1. The Ising prior provides superior performance in terms of pure misclassification rates as visible in the table above. Overall, the simulation reveals the Gaussian prior as the weaker of the two priors used for this application, and the Ising as the most robust. Therefore, from this example it can be concluded that in the case of the Ising model the numerical approximation to the normalizing constant appears highly accurate in the empirical work done by the author.

Wei and Pan[62] proposed comparing discrete and Gaussian Markov random fields (MRF) and understand how these methods fare to analyze genomic data in a gene network modelling. Because the genomic data are in the usual format of DNA microarray expression data, it is technically possible to apply any of many existing statistical methods of detecting differentially expressed genes to binding data. The author[62] incorporate gene networks as prior

	% Voxels misclassified	
Noise level (c)	Ising prior	Gaussian prior
0.5	**0.095**	0.109
1	**0.642**	0.698
2	**1.748**	1.781
	% Foreground voxels misclassified	
Noise level (c)	Ising prior	Gaussian prior
0.5	**2.415**	2.814
1	**16.754**	18.271
2	**45.754**	46.695

Table 1: Average Performance Metrics for 50 Replicates of the fMRI Simulation Based on the Ising and Gaussian prior. The metrics corresponding to the best performing prior are in bold.

biological knowledge into statistical modelling of micro array data to maximize the power for biological discoveries. Firstly, a comparison of inferences was done by taking a close look on the conditional distributions of each prior given the data and all other parameters in the model. This also helped in throwing light on the model complexities. The conditional probability of Gaussian MRF was given as:

$$Pr(T_i = 1|\boldsymbol{z}, \theta, x_{i0}, x_{i1}) = \frac{1}{1+\exp(x_{i0}-x_{i1})\phi(z_i;\mu_0,\sigma_0^2)/\phi(z_i;\mu_1,\sigma_1^2)}, \tag{30}$$

where T_i is the state of gene i, z_i is the test statistic measuring the relative abundance of the Transcription factors (TF), $\theta = (\mu_0, \mu_1, \sigma_0, \sigma_1)$, $\phi(\mu, \sigma^2)$ is the density function for a normal distribution with mean μ and variance σ^2, $\boldsymbol{x}$ is the G-dimensional latent vectors distributed according to an intrinsic Gaussian conditional auto-regression model. The conditional probability of Discrete MRF was given as:

$$Pr(T_i = 1|\boldsymbol{z}, \theta, T_{\partial i}, \Phi) = \frac{1}{1+(1/\exp[\gamma+\beta\{n_i(1)-n_i(0)\}/m_i])\phi(z_i;\mu_0,\sigma_0^2)/\phi(z_i;\mu_1,\sigma_1^2)}, \tag{31}$$

where $\Phi = (\gamma, \beta)$, γ and $\beta > 0$ are arbitrary real numbers, ∂i represents the (direct) neighbors of gene i and $n_i(j)$ is the number of gene $i^{'}$s neighbors having states j for $j = 0, 1$ and m_i is the number of gene $i^{'}$s neighbors. The attraction parameter β corresponds to the spatial interaction strength in the Discrete MRF, i.e the tendency of sharing the same state for neighboring genes.

Although in practice, inferences are based on the marginal posterior probability $Pr(T_i|\boldsymbol{z})$, the above conditional posterior probabilities provide a unique perspective to compare the two prior models. First, the (conditional) posterior probability of being a target is jointly determined by the prior probability ratio and the data, i.e. the likelihood ratio, in both the models. This sheds light on that misspecified prior distributions, e.g. due to incomplete gene networks, may not have a large influence on the posterior probability if the data likelihood ratio is large. Second, the GMRF is more richly parameterized by introducing thousands of additional parameters ($x_{ij}s$), compared with the DMRF. However, these additional parameters are not treated as independent fixed effects but are linked by the adopted hierarchical structure of the GMRFs,which leads to borrowing information among the parameters via shrinkage (Carlin and Louis[63]). The extent of the shrinkage among $x_{ij}s$ is controlled by σ_{C0} and σ_{C1}. For example, when they are both 0, x_{ij} is a constant.

The author[62] in order to compare the performances of both priors, downloaded the GCN4 CHIP–chip data from the web site of Lee et al.[64] where Binding ratios and p-values for 6270 yeast genes were available. The ROC curves were constructed for both the methods on the basis of the positive and negative control sets. When the specificity ranged from 0.9 to 0.4, the Bayesian DMRF had higher sensitivities compared with that of the GMRFMM with the zero constraint. It was also shown by the author[62] that DMRF encouraged less clustering as compared to GMRF with zero constraint. To compare the methods further, particularly their robustness to misspecified gene networks, [62] conducted a simulation study that mimicked real data: i.e used the same gene network as used for the real data and used data-generating distributions that were similar to those fitted to the real data. They applied the true network to each of the 20 simulated data sets and constructed the ROC curves and found that GMRF with zero constraint had similar sensitivity at a high specificity.

Kappes et al[65] performed a comparative study of Inference Techniques for Discrete Markov Random fields (DMRF). OpenGM 2, a C++ library for discrete graphical models has been used for the purpose. The author discusses Pixel-based-Models among others where each pixel in a 2D lattice is a variable in the model and has four nearest neighbors. The study incorporated numerous models, however to name a few $mrf - stereo$, $mrf - inpainting$, and $mrf - photomontage$ from [66] with three, two and two instances, respectively. All these models had a 4 neighborhood assumption. The author then evaluated a large number of different inference methods such as : (i) combinatorial-methods (ii) linear- programming-methods (iii) move-making methods (iv) message passing methods. In all these inference methods, algorithms based on Monte Carlo simulation were not considered. Different inference algorithms (from three methods mentioned above) were applied in Pixel models. The findings for Stereo Matching (mrf-stereo), Inpainting (mrf-inpainting) and Photomontage (mrf-photomontage) has been summarised.

StereoMatching (mrf-stereo): The author[65] considers three instances of a graphical model for the stereo matching problem in vision. Results are shown in Tab. 2. On average, TRWS-LF2 algorithm from Linear programming method and CombiLP algorithm from Combinatorial methods afford the best solutions. FastPD algorithm from Move-making method is the fastest algorithm. Solutions obtained by BPS which is a Message-Passing method are better in terms of the two-pixel accuracy (PA2) than solutions with lower objective value. Storing the functions of a graphical model explicitly, as value tables, instead of as implicit functions, slows algorithms down, as can be seen for TRWS. It can be seen from these results that two instances were solved to optimality. For the two instances for which optimal solutions were obtained, suboptimal approximations that were obtained significantly faster are not significant worse in terms of the two pixel accuracy (PA2), i.e. the number of pixels whose disparity error is less than or equal to two. On average, the solution obtained by BPS is 2% better in terms of the two-pixel accuracy.

algorithm	runtime	value	bound	PA2
FastPD	3.34 sec	1614255.00	-∞	0.6828
mrf-α-Exp-TL	10.92 sec	1616845.00	-∞	0.6823
mrf-$\alpha\beta$-Swap-TL	10.14 sec	1631675.00	-∞	0.6832
ogm-LF-2	366.02 sec	7396373.00	-∞	0.3491
ogm-TRWS-LF2	439.75 sec	**1587043.67**	1584746.53	0.6803
mrf-LBP-TL	287.62 sec	1633343.00	-∞	0.6804
mrf-BPS-TL	238.70 sec	1738696.00	-∞	**0.7051**
mrf-TRWS-TAB	1432.57 sec	1587681.33	1584725.98	0.6806
mrf-TRWS-TL	227.67 sec	1587928.67	1584746.53	0.6803
ogm-CombiLP	969.33 sec	1587560.67	1584724.04	0.6809

Table 2: mrf-stereo (3 instances)

Inpainting (mrf-inpainting): The author[65] considers two instances of a graphical model for image inpainting. In these instances, every variable can attain 256 labels. Results are shown in Tab. 3. It can be seen from these results that TRWS outperforms move-making methods such as α-expansion and FastPD. The best result is obtained by taking the solution provided by TRWS as the starting point for a local search by lazy flipping. While FastPD and α-expansion converge faster than TRWS, their solution is significantly worse in terms of the objective value and also in terms of the mean color error (CE).

algorithm	runtime	value	bound	CE
FastPD	8.47 sec	32939430.00	-∞	14.7
mrf-α-Exp-TL	54.21 sec	27346899.00	$-\infty$	11.3
mrf-$\alpha\beta$-Swap-TL	111.13 sec	27154283.50	$-\infty$	12.0
ogm-TRWS-LF2	3009.52 sec	**26463829.00**	**26462450.59**	10.9
mrf-LBP-TL	666.19 sec	26597364.50	-∞	**10.5**
mrf-BPS-TL	644.15 sec	26612532.50	$-\infty$	12.0
mrf-TRWS-TL	614.05 sec	26464865.00	**26462450.59**	10.9
ogm-CombiLP	49672.26 sec	26467926.00	26461874.39	10.9

Table 3: mrf-inpainting (2 instances)

Photomontage (mrf-photomontage). The author[65] now consider two instances of graphical models for photomontage. Results are shown in Tab. 4. It can be seen from these results that move-making algorithms such as α-expansion outperform algorithms based on linear programming relaxations such as TRWS. The reason behind this observation is explained by the fact that the second-order factors are more discriminative than the first-order factors. Therefore, the LP relaxation is loose and good primal solution is difficult.

Overall, in order to model spatial-temporal interactions in the thermographic brain imaging data, the semi parametric regression in the generalized linear mixed model approach seems to be an good choice. By discussing two different types of Markov random fields, Discrete Markov Random fields (DMRF) has been shown to be a better method in spatial smoothing.

algorithm	runtime	value	bound
mrf-α-Exp-TAB	**7.54** sec	**168284.00**	-∞
mrf-$\alpha\beta$-Swap-TAB	8.42 sec	200445.50	$-\infty$
ogm-TRWS-LF2	390.34 sec	735193.00	**166827.12**
mrf-LBP-TAB	686.61 sec	438611.00	-∞
mrf-BPS-TAB	167.49 sec	2217579.50	$-\infty$
mrf-TRWS-TAB	172.20 sec	1243144.00	**166827.07**

Table 4: mrf-photomontage (2 instances)

References

[1] Weisberg, S. *Applied Linear Regression*, 2nd Ed., J. Wiley & Sons, Inc., New York 1985

[2] Draper, Norman Richard., and Harry Smith. *Applied Regression Analysis*. 3rd ed. New York ; Chichester: John Wiley, 1998.

[3] D. Ruppert, M. P. Wand, and R. J. Carroll, *Semiparametric Regression*, Cambridge University Press, New York (2003)

[4] C. DeBoor, *A practical guide to splines*, Springer-Verlag, New York (1978).

[5] Eilers, P.H.C. and B.D. Marx, *Flexible Smoothing with B-Splines and Penalties*. Statistical Science, 11(2):89-121,1996

[6] Brian D Marx and Paul H.C Eilers, *Multidimensional Penalized Signal Regression*, Technometrics 47(1)·February 2005

[7] Paul H.C. Eilers, Iain D. Currie and Maria Durbán, *Fast and compact smoothing on large multidimensional grids*, Computational Statistics & Data Analysis 50 (2006) 61–76

[8] William R. Schucany, *Kernel Smoothers: An Overview of Curve Estimators for the First Graduate Course in Nonparametric Statistics*, Statistical Science 2004, Vol. 19, No. 4, 663–675

[9] J. Duchon, 1976, *Splines minimizing rotation invariant semi-norms in Sobolev spaces.* pp 85–100, In: Constructive Theory of Functions of Several Variables, Oberwolfach 1976, W. Schempp and K. Zeller, eds., Lecture Notes in Math., Vol. 571, Springer, Berlin, 1977

[10] Fred L. Bookstein, *Principal Warps: Thin-Plate Splines and the Decomposition of Deformations*,IEEE Transactions on Pattern Analysis and Machine Intelligence. VOL. II . No. 6. June 1989

[11] Simon N.Wood, *Thin plate regression splines*, J. R. Statist. Soc. B (2003) 65, Part 1, pp. 95–114

[12] Iain D. Currie, Maria Durbán and Paul H.C. Eilers(2006), *Generalized linear array models with applications to multidimensional smoothing*, Journal of the Royal Statistical Society, Series B, 68, 1–22.

[13] Tim Ramsay, *Spline smoothing over difficult regions*, J. R. Statist. Soc. B (2002) 64, Part 2, pp. 307–319

[14] Simon N.Wood, Mark V. Bravington and Sharon L. Hedley, *Soap film smoothing*, J. R. Statist. Soc. B (2008) 70, Part 5, pp. 931–955

[15] Haonan Wang and M. Giovanna Ranalli, *Low-Rank Smoothing Splines on Complicated Domains*, Biometrics 63, 209–217, March 2007

[16] David L. Miller and Simon N. Wood, *Finite area smoothing with generalized distance splines*, Environ Ecol Stat (2014) 21:715–731

[17] Simon N. Wood(2004), *Stable and Efficient Multiple Smoothing Parameter Estimation for Generalized Additive Models*, Journal of the American Statistical Association, 99:467, 673-686

[18] E. E. Kammann and M. P.Wand, *Geoadditive models*, Appl. Statist. (2003) 52, Part 1, pp. 1–18

[19] Dae-Jin Lee and Maria Durban(2011), *P-spline ANOVA-type interaction models for spatio-temporal smoothing*, Statistical Modelling 2011; 11(1): 49–69

[20] Ludwig Fahrmeir, Thomas Kneib and Stefan Lang(2004), *Penalized Structured Additive Regression for Space-time Data: A Bayesian Perspective*, Statistica Sinica 14(2004), 731-761

[21] Xihong Lin, Daowen Zhang(1999), *Inference in generalized additive mixed models by using smoothing splines*, J. R. Statist. Soc. B(1999)61

[22] Fahrmeir, L. and Knorr-Held, L. (2000), *Dynamic and semiparametric models. In Smoothing and Regression: Approaches, Computation and Application (Edited by M. Schimek).*, Wiley, New York

[23] Fahrmeir, L. and Lang, S. (2001a). *Bayesian inference for generalized additive mixed models based on Markov random field priors*, J. Roy. Statist. Soc. Ser. C 50, 201-220.

[24] Fahrmeir, L. and Lang, S. (2001b). *Bayesian semiparametric regression analysis of multicategorical time-space data.* Ann. Inst. Statist. Math. 53, 10-30.

[25] Fahrmeir, L., Lang, S., Wolff, J. and Bender, S. (2003). *Semiparametric Bayesian time-space analysis of unemployment duration.* J. German Statist. Soc. 87, 281-307.

[26] Fahrmeir, L. and Tutz, G. (2001). *Multivariate Statistical Modelling based on Generalized Linear Models.*, Springer-Verlag, New York.

[27] Stefan Lang and Ludwig Fahrmeir, *Bayesian generalized additive mixed models. A simulation study*, Sonderforschungsbereich 386, Paper 230 (2001)

[28] Andreas Brezger, Thomas Kneib and Stefan Lang, *BayesX: Analyzing Bayesian Structured Additive Regression Models.* Journal of Statistical Software, September 2005, Volume 14, Issue 11

[29] Clemens Biller, *Adaptive Bayesian Regression Splines in Semiparametric Generalized Linear Models.* Sonderforschungsbereich 386, Paper 133 (1998)

[30] Dani Gamerman, *Sampling from the posterior distribution in generalized linear mixed models.* Statistics and Computing (1997) 7, 57±68

[31] Andreas Brezger and Stefan Lang, *Generalized structured additive regression based on Bayesian P-splines.* Sonderforschungsbereich 386, Paper 321 (2003)

[32] Julian Besag, *Spatial Interaction and the Statistical Analysis of Lattice Systems*, Journal of the Royal Statistical Society. Series B (Methodological), Vol. 36, No. 2. (1974), pp. 192-236.

[33] Kindermann, R and Snell,J .L. (1980) *Markov Random Fields and their Applications* (Providence, RI, American Mathematical Society).

[34] Derin, H. & Elliot, H. (1987) *Modeling and segmentation of noisy and textured images using Gibbs random fields*, IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-9, pp. 39-55.

[35] G. R. Cross and A. K. Jain, *Markov random field texture models*, IEEE Trans. Pattern Anal. Machine Intell., vol. PAMI-5, pp. 25-39, 1983.

[36] S. Geman and D. Geman, *Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images*, IEEE Trans. Pattern Anal. Machine Intell., vol. PAMI-6, pp. 721-741, Nov. 1984.

[37] F. S. Cohen and D. B. Cooper, *Real time textured image segmentation based on noncausal Markovian random field models*. in Proc. SPIE Conf Intell. Robots, Cambridge, MA, Nov. 1983.

[38] H. Derin and W. S. Cole, *Segmentation of textured images using Gibbs random fields*. Computer Vision, Graphics, Image Processing, vol. 35, pp. 72-98, 1986.

[39] Chellappa R (1985) *Two-dimensional discrete Gaussian Markov random field models for image processing* in: L.N. Kanal& A. Rosenfeld(Eds) Progress in Pattern Recognition 2, pp. 79-112 (Amsterdam, North-Holland).

[40] Besag J (1975) *Statistical analysis of non-lattice data*, The Statistician, 24, pp. 179-195.

[41] Besag, J. (1977) *Efficiency of pseudo-likelihood estimation for simple Gaussian fields*, Biometrika, 64, pp. 616-618.

[42] Manjunath and Chellappa (1990), *Stochastic and Deterministic Networks for Texture Segmentation*. IEEE Transactions on Acoustics. Speech. and Signal Processing. Vol. 3X. No. 6. June 1990

[43] Abend, K., Harley, T. J. and Kanal, L. N. (1965), *Classification of binary random patterns.* IEEE Trans. Inform. Theory IT-11, 538-544.

[44] J.M. Hammersley and P. Clifford. *Markov fields on finite graphs and lattices.* University of California, Berkeley

[45] Julian Besag, *On the Statistical Analysis of Dirty Pictures.* J. R. Statist. Soc. B (1986) 48, No.3, pp. 259-302

[46] Thrasyvoulos N. Pappas, *An Adaptive Clustering Algorithm for Image Segmentation.* IEEE Transactions on Signal Processing VOL 10 NO 1 April 1992

[47] A. M. Gorbach, J. Heiss, C. Kufta, S. Sato, P. Fedio, W. A. Kammerer, J. Solomon, and E. H. Oldfield. *Intraoperative infrared functional imaging of human brain.* Ann. of Neurol., 54:297–13, 2003

[48] A. M. Gorbach, J. D. Heiss, L. Kopylev, and E. H. Oldfield. *Intraoperative infrared imaging of brain tumors.* J Neurosurg, 101(6):960–9, 2004.

[49] I. A. Shevelev, E. N. Tsicalov, A. M. Gorbach, K. P. Budko, and G. A. Sharaev. *Thermoimaging of the brain.* J. Neurosci. Methods, 46:49–9, 1993.

[50] Nico Hoffmann, Yordan Radev, Edmund Koch, Gerald Steiner, Matthias Kirsch, and Uwe Petersohn. *Intraoperative identification of the sensory cortex by semiparametric regression of time-resolved thermography..*

[51] Nelder, John; Wedderburn, Robert (1972). *Generalized Linear Models.* Journal of the Royal Statistical Society. Series A (General). Blackwell Publishing. 135 (3): 370–384.

[52] C. Sidney Burrus. *Iterative Reweighted Least Squares.* http://cnx.org/content/m45285/1.12/

[53] Jerome H. Friedman, Bernard W. Silverman(1989). *Flexible Parsimonious Smoothing and Additive Modeling.* Technometrics, Vol 31, No.1. (Feb 1989), pp. 3-21

[54] Finbarr O'Sullivan;Brian S. Yandell;William J. Raynor(1986). *Automatic Smoothing of Regression functions in Generalized Linear Models*. Journal of American Statistical Association, Vol.81, No.393. (Mar 1986),pp-96-103

[55] Usha S. Govindarajulu, Elizabeth J. Malloy, Bhaswati Ganguli, Donna Spiegelman, Ellen A. Eisen (2009). *The Comparison of Alternative Smoothing Methods for Fitting Non-Linear Exposure-Response Relationships with Cox Models in a Simulation Study*

[56] Hastie, T. J.; Tibshirani, R. J. (1990). *Generalized Additive Models.* Chapman & Hall/CRC. ISBN 978-0-412-34390-2.

[57] Maria Durban, Iain Currie and Paul Eilers. *Multidimensional P-spline Mixed Models: A unified approach to smoothing on large grids.*

[58] Mahalanobis, Prasanta Chandra (1936). *On the generalised distance in statistics.* Proceedings of the National Institute of Sciences of India. 2 (1): 49–55. Retrieved 2016-09-27.

[59] Grace Wahba (1990). *Spline Models for Observational Data.* Society for Industrial and Applied Mathematics Philadelphia,Pennsylvania 1990

[60] Daniel Smith & Michael Smith (2006). *Estimation of Binary Markov Random Fields Using Markov chain Monte Carlo.* Journal of Computational and Graphical Statistics, 15:1, 207-227, DOI: 10.1198/106186006X97817

[61] Hurn, M. (1998), *Confocal Fluorescence Microscopy of Leaf Cells: An Application of Bayesian Image Analysis.* Applied Statistics, 47, 361–377.

[62] Peng Wei & Wei Pan (2008). *Network-based genomic discovery: application and comparison of Markov random-field models.* Appl. Statist. (2010) 59, Part 1, pp. 105–125

[63] Carlin, B. P. and Louis, T. A. (2000) *Bayes and Empirical Bayes Methods for Data Analysis*, 2nd edn. New York: Chapman and Hall–CRC.

[64] Lee, T. I., Rinaldi,N. J., Robert, F., Odom,D. T., Bar-Joseph, Z., Gerber,G. K., Hannett,N. M., Harbison, C. T., Thompson, C. M., Simon, I., Zeitlinger, J., Jennings, E. G., Murray, H. L., Gordon, D. B., Ren,

B.,Wyrick, J. J., Tagne, J. B., Volkert, T. L., Fraenkel, E., Gifford, D. K. and Young,R. A. (2002) *Transcriptional regulatory networks in Saccharomyces cerevisiae.* Science, 298, 799–804.

[65] Jörg H. Kappes and Björn Andres and Fred A. Hamprecht and Christoph Schnörr and Sebastian Nowozin and Dhruv Batra and Sungwoong Kim and Bernhard X. Kausler and Thorben Kröger and Jan Lellmann and Nikos Komodakis and Bogdan Savchynskyy and Carsten Rother. *A Comparative Study of Modern Inference Techniques for Structured Discrete Energy Minimization Problems.* CoRR, abs/1404.0533, 2014, http://arxiv.org/abs/1404.0533

[66] R. Szeliski, R. Zabih, D. Scharstein, O. Veksler, V. Kolmogorov, A. Agarwala, M. Tappen, and C. Rother. *A comparative study of energy minimization methods for Markov random fields with smoothness-based priors.* IEEE PAMI, 30(6):1068–1080, 2008.

[67] V. Kolmogorov. *Convergent tree-reweighted message passing for energy minimization.* PAMI, 28(10):1568–1583, 2006.

[68] Daniela M. Witten, Robert Tibshirani and Trevor Hastie (2009). *A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis.* Biostatistics (2009), 10, 3, pp. 515–534 doi:10.1093/biostatistics/kxp008

[69] Andrew Y. Ng. *Feature selection, L1 vs L2 regularization, and rotational invariance.* Computer Science Department, Stanford University, Stanford, CA 94305, USA