PFG 2013 / 4, 0255 – 0267
Stuttgart, August 2013

Article

Graphical Models in Geodesy and Photogrammetry

WOLFGANG FÖRSTNER, Bonn

Keywords: graphical models, Bayesian nets, Markov random fields, conditional random fields, geodetic networks, bundle adjustment, iterative conditional modes, Gauss-Seidel iteration

Summary: The paper gives an introduction into graphical models and their use in specifying stochastic models in geodesy and photogrammetry. Basic task in adjustment theory can intuitively be described and analysed using graphical models. The paper shows that geodetic networks and bundle adjustments can be interpreted as graphical models, both as Bayesian networks or as conditional random fields. Especially hidden Markov random fields and conditional random fields are demonstrated to be versatile models for parameter estimation and classification.

Zusammenfassung: *Graphische Modelle in Geodäsie und Photogrammetrie*. Der Beitrag gibt eine Einführung in Graphische Modelle und ihren Einsatz zur Erstellung probabilistischer Modelle in der Geodäsie und der Photogrammetrie. Grundaufgaben der Ausgleichungsrechnung lassen sich intuitiv beschreiben und analysieren. Der Beitrag zeigt, wie geodätische/photogrammetrische Netze als Bayesnetze oder Markoff-Zufallsfelder interpretiert werden können. Besonders bedingte Zufallsfelder erweisen sich als flexibel für die Modellierung und die Optimierung von Parameterschätz- und Klassifikationsaufgaben.

1 Introduction

For more than 50 years geodetic and photogrammetric networks are classical tools for point positioning and orientation determination. They are characterized by a sparse link between observed data, image coordinates, distances, angles, height differences, or GPS-coordinates, on one hand and unknown parameters, mostly coordinates but also orientation parameters or additional parameters on the other hand, to capture various systematic effects. The work horse for determining the unknown parameters in a statistically optimal manner is the classical adjustment theory, including its Bayesian variants which allows for including prior information or sequential estimation.

In the last decades so-called graphical models, especially Markov random fields, have found their way into photogrammetric research, mainly for image interpretation. In contrast to simple pixel-wise or image region-wise classifiers these models allow for statistical modelling the spatial neighbourhood between pixels or image regions, and lead to an increase in classification performance.

The tools for finding optimal classifications based on these graphical models are in no way related to methods for solving large equation systems. Even more, in most cases only approximate solutions can be found and the statistical properties of the results are difficult to characterize.

Therefore the question arises: How is the relation between the current methodology and the new one? Do they address different problems? Is there an overlap? Is the current methodology a special case of the new one? The answer is clear: Graphical models are a real generalization of the well-known tools from statistics and adjustment theory, in the way they are trained and used in geodesy and photogrammetry. If a new methodology comes up, which claims to be a true generalization of the current one, it has to prove (a) that the current methods can be derived from the new ones by specialization and (b) that there is a relevant potential for successfully solving new problems, which cannot

www.schweizerbart.de
DOI: 10.1127/1432-8364/2013/0175
1432-8364/13/0175 $ 3.25

be tackled by the current methodology. The paper mainly addresses the first point. The second point is sketched and proven in some of this issue's articles.

Graphical models are probabilistic models on graphs, the nodes representing single or aggregated random variables, and the edges representing probabilistic relationships between the random variables. In case these relations are directed, the networks fall into the category of Bayesian networks, otherwise they are so-called Markov random fields. In the above mentioned application of interpreting images, the random variables in the graphical models are discrete, explaining the conceptually different procedures for finding optimal solutions. However, graphical models are not restricted to model and analyse situations with discrete random variables, i.e. classification tasks, but also can handle problems with continuous variables. A prominent example is the classical Kalman filter, see Fig. 1. It is a graphical model. It con-

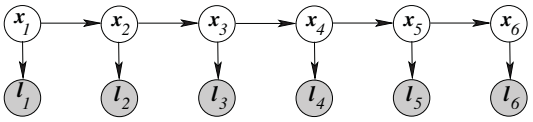

Fig. 1: A hidden Markov-chain as a special continuous Bayesian network used as basic model in Kalman filtering. The joint probability of all nodes can be factorized into $p(\boldsymbol{l}, \boldsymbol{x}) = p(x_1) \prod_{t=2}^{6} p(x_t|x_{t-1}) \prod_{t_1}^{6} p(l_t|x_t)$, each factor depending only on one or two variables. This eases learning and reasoning. The naming *Markov-chain* indicates, that (1) the time series follows the Markov-property, i.e. future states do only depend on the present, not on past states, and (2) the structure of the underlying graph of the unknown states (white) is a chain, i.e. only has nodes with maximum two neighbours.

tains two types of nodes $\boldsymbol{x}_t$ and $\boldsymbol{l}_t$. The nodes $\boldsymbol{x}_t$ represent the unknown state, shown in white, varying over time, building a so-called hidden Markov chain, as the new state only depends on the previous state, and not on older states. The nodes $\boldsymbol{l}_t$ represent the observations at time t. Here, we assume observed values are available, therefore the nodes are shown in grey.

The methods for optimal estimation, prediction and filtering of such a Bayesian network are well known, not only in geodesy and photogrammetry.

This paper wants to uncover more correspondences between graphical models and geodetic and photogrammetric parameter estimation problems, namely showing the close relation between geodetic networks and photogrammetric blocks and so-called conditional random fields, which have shown their power in image interpretation during the last years.

The paper does not go into the details of graphical models, which are documented in quite a number of lectures, e.g. BILMES (2000), MURPHY (1998) and books, e.g. PEARL (1988b), LI (2000), BISHOP (2006), WINKLER (2006).

The paper is organized as follows. We first give a short introduction into graphical models, relating the concepts to basic statistical tasks in geodetic education. Using several examples, we show the close relation between adjustment theory and Bayesian networks. Using a simple four-node network we on one hand demonstrate the versatility of graphical models to describe the probabilistic models for geodetic networks and for image interpretation, on the other hand uncover the intimate link between independence relations in Markov random fields and the sparsity of classical normal equation matrices.

2 Graphical Models

2.1 *Motivation*

The complexity of probabilistic models increases exponentially with the number of variables if no structure is imposed. Take as an example a small binary image of 326×119 pixels as shown in Fig. 2. In order to describe the joint

Fig. 2: Binary image with 326×119 pixels.

probability of all $N = 326 \times 119 = 38794$ pixels, one would need an enormous number of $2^N - 1 = 2^{38794} - 1 \approx 10^{11678}$ probabilities $P(x_1, ..., x_{38794})$. The number of high resolution colour images is even larger. In order to get an impression of the size of this large number

10^{11678}, one can relate it either to the comparably really microscopic number of 10^{78} atoms in the universe, or to the a bit larger number 10^{130} of possible images taken at any micrometer within the universe at any microsecond of its estimated lifetime in one of 100 million directions.

Probabilistic models with a large number of variables usually reveal an internal structure, which therefore needs to be exploited. The structure on one hand results from the object modelled. As an example take the model of time-dependent processes, which refers to the states in a sequence of points in time, see Fig. 1, or take the stochastic model of human body configuration in image analysis, which is related to the pose of connected limbs.

As geodesists and photogrammetrists we usually work with a simplified assumption about the distribution, a Gaussian distribution. As a consequence the number of parameters necessary to specify the distribution only increases quadratically with the number of parameters, which for a million variables would require the specification of $\approx 10^{12}$ entries in the covariance matrix. The ability to approximate covariance matrices using covariance functions of large point clouds suggests that in some applications much less parameters are necessary to arrive at an adequate stochastical model. Another geodetic approach to reduce the complexity of a stochastical model is the concept of primary errors, which allows to explain correlations in high dimensional covariance matrices using only a few causing effects.

2.2 *Definition and Types*

A graphical model is a probabilistic model where the dependence structure between the variables is described by a graph $G(\mathcal{N}, \mathcal{E})$. The nodes $n \in \mathcal{N}$ represent stochastical variables, the edges $e = (n_i, n_j) \in \mathcal{E}$ represent probabilistic relations.

A node may represent a single random variable, say $\underline{x}$, following some distribution, e.g. specified by $\underline{x} \sim p_x(x)$. It may be a set or vector $\underline{\boldsymbol{x}}$ of variables or even a more complex structure of random variables. Random variables are indicated by an underscore. The index x in p_x indicates the name of the variable the density function refers to, which we omit, in case there is no confusion. The graphical presentation shows the name, say x, and possibly the type of the random variable, e.g. $\boldsymbol{x}$ if it is a vector. The random variables may be discrete or continuous, or mixed. The distribution of the variables may be completely unknown, partially or fully known. This flexibility of course requires to specify the content of the nodes in the legend of the graphical model. In the following we will assume the nodes to either represent a vector of continuous random variables, e.g. standing for measurements or parameters useful in parameter estimation problems, or a discrete random variable, e.g. standing for a class name out of a given set of possible classes useful in classification or interpretation problems.

The variable of a node may be either unknown, then the node is drawn with a white background, or it is observed, equivalent to having a sample value of the underlying distribution, then it is drawn with a grey background.

The edges of the graph are either directed or undirected. Directed edges are used, in case one wants to specify conditional probabilities.

Fig. 3 shows several graphical models with two nodes. For directed networks we show various variants with observed nodes and repetitions.

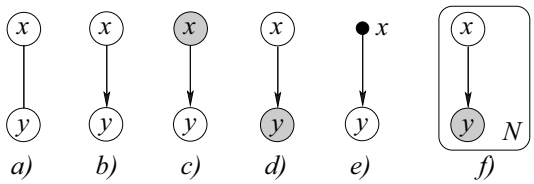

Fig. 3: Graphical models with two nodes. White nodes: unobserved. Grey nodes: observed. Black node: fixed value. Undirected edge: $p(x, y)$ is part of the model. Directed edge: $p(y|x)$ is part of the model. The rounded box indicates that the two-node network exists N-times.

The essential idea of graphical models is that the joint probability distribution $p(\boldsymbol{x}) = p(x_1, x_2, ..., x_n, ..., x_N)$ can be written as a product

$$p(x_1, x_2, ..., x_n, ..., x_N) = \frac{1}{Z} \prod_i f_i(\mathcal{X}_i) \quad (1)$$

of functions f_i of small subsets $\mathcal{X}_i = \{x_{i_1}, ..., x_{i_{N_i}}\}$ of variables, and these sets $\mathcal{X}_i$ can be seen in the graph. The functions either

result from some statistical knowledge or just can be chosen such that large values of f_i support the joint probability $p(\boldsymbol{x})$. The constant Z, the so-called partition function, sometimes is necessary to guarantee that the resulting probability density fulfills $\int p(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = 1$.

In the following we show that classical models in geodesy and photogrammetry can be modelled using graphical models.

2.3 *Bayesian Nets*

2.3.1 The model for collocation

In time series often a noisy signal is observed, and the original signal is to be recovered. The problem is known as collocation in geodesy and photogrammetry, see MORITZ (1978, section 4), KRAUS (1972). The signal, say $\underline{\boldsymbol{x}}$, is assumed to be a random vector. Its covariance structure usually is described using covariance functions, a modelling tool also having found its place in pattern recognition via so-called Gaussian processes, see RASMUSSEN & WILLIAMS (2005). The signal is contaminated by noise, say $\underline{\boldsymbol{n}}$. The noise is assumed to be independent of the signal and again, in case it is vector valued, may be correlated. The resulting observed signal, say $\boldsymbol{y}$, is a function of the unknown random signal $\boldsymbol{x}$ and the noise $\boldsymbol{n}$, say by addition. The Bayesian network in Fig. 4 is a graphical model for this situation. Observe, the model only specifies the principal relationship. The joint probability of all three variables can

Fig. 4: A model for filtering a noisy signal. The signal $\boldsymbol{x}$ is contaminated by noise $\boldsymbol{n}$, independent on $\boldsymbol{x}$. The contaminated signal is $\boldsymbol{y}$, depending on both $\boldsymbol{x}$ and $\boldsymbol{n}$. Left: model prior to an observation. Right: The situation where the contaminated signal is observed: In case certain characteristics of the signal $\boldsymbol{x}$ and the noise $\boldsymbol{n}$ are known, both can be recovered from the observed signal $\boldsymbol{y}$.

easily derived from the graph. The nodes with no parents, $\boldsymbol{x}$ and $\boldsymbol{n}$ are assumed to be independent and follow some distribution, the nodes $\boldsymbol{y}$ is depending on these two nodes, and the conditional probability $p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{n})$ is assumed to be part of the model. Thus, the joint probability of all three variables is

$$p(\boldsymbol{x},\boldsymbol{n},\boldsymbol{y}) = p(\boldsymbol{x})\; p(\boldsymbol{n})\; p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{n}) \qquad (2)$$

indicating no specific density $p(\boldsymbol{x})$ and $p(\boldsymbol{n})$ for the independent variables nor for the conditional probability $p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{n})$. Thus, when specifying the model, there is no need to specify the densities.

In case all variables are zero mean Gaussian distributed and the contamination model is additive, i.e. $\underline{\boldsymbol{y}} = \underline{\boldsymbol{x}} + \underline{\boldsymbol{n}}$, the covariance matrix of the joint vector $[\underline{\boldsymbol{x}}, \underline{\boldsymbol{n}}, \underline{\boldsymbol{y}}]^\mathsf{T}$ would be

$$\mathbb{D}\left(\begin{bmatrix} \underline{\boldsymbol{x}} \\ \underline{\boldsymbol{n}} \\ \underline{\boldsymbol{y}} \end{bmatrix}\right) = \begin{bmatrix} \Sigma_{xx} & 0 & \Sigma_{xx} \\ 0 & \Sigma_{nn} & \Sigma_{nn} \\ \Sigma_{xx} & \Sigma_{nn} & \Sigma_{xx}+\Sigma_{nn} \end{bmatrix} \qquad (3)$$

which follows from variance propagation. Given an observed value $\boldsymbol{y}$ for the contaminated signal, using Bayesian theorem in the form $p(\boldsymbol{x},\boldsymbol{n}|\boldsymbol{y}) = p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{n})p(\boldsymbol{x})p(\boldsymbol{n})/p(\boldsymbol{y})$ one in the general setting can derive the density $p(\boldsymbol{x}|\boldsymbol{y})$ of the signal given $\boldsymbol{y}$ and the density $p(\boldsymbol{n}|\boldsymbol{y})$ of the noise given $\boldsymbol{y}$.

In the case of normally distributed variables, with

$$\Sigma_{yy} = \Sigma_{xx} + \Sigma_{nn} \qquad (4)$$

we obtain the classical result

$$\mathbb{E}\left(\begin{bmatrix} \underline{\boldsymbol{x}}|\boldsymbol{y} \\ \underline{\boldsymbol{n}}|\boldsymbol{y} \end{bmatrix}\right) = \begin{bmatrix} \Sigma_{xx} \\ \Sigma_{nn} \end{bmatrix} \Sigma_{yy}^{-1}\boldsymbol{y} \qquad (5)$$

and

$$\mathbb{D}\left(\begin{bmatrix} \underline{\boldsymbol{x}}|\boldsymbol{y} \\ \underline{\boldsymbol{n}}|\boldsymbol{y} \end{bmatrix}\right) = \begin{bmatrix} \Sigma_{xx} \\ \Sigma_{nn} \end{bmatrix} \Sigma_{yy}^{-1}\,[\Sigma_{xx}|\Sigma_{nn}] \qquad (6)$$

indicating the derived signal and noise variables $\underline{\boldsymbol{x}}|\boldsymbol{y}$ and $\underline{\boldsymbol{n}}|\boldsymbol{y}$ are 100 % correlated.

This result can be directly transferred into a rule interpreting the independence relations in a Bayesian network. In a three-node network of the type in Fig. 4 left, the parent nodes $\boldsymbol{x}$ and $\boldsymbol{n}$ are independent of $\boldsymbol{y}$, whereas in the right network the parent nodes $\boldsymbol{x}$ and $\boldsymbol{n}$ are conditionally dependent given $\boldsymbol{y}$. This result is independent of the type of density functions involved. The conditional dependency can be used to infer from a subsequent observation of one of the

two causing variables to the other variable – a special case of PEARL's (1988a) so-called *explaining away*. Similar interpretation rules exist for chains and can be used to analyse more complex networks.

2.3.2 The concept of primary errors

A well-established geodetic principle to obtain a probabilistic model for the observations with only a few parameters is the concept of primary errors (PELZER 1974): Correlations in the high-dimensional covariance matrix of observations are explained by a few, say K unknown effects $\underline{p}_k$, which, besides individual random perturbations $\underline{e}$, influence all N variables $\underline{l}_n$. This also is the basis for modelling systematic image errors using additional parameters in bundle adjustment. With the, in general unknown, mean values $\underline{y}_n$ the general model reads as

$$p(l_n|y_n, \boldsymbol{p}) = \prod_n p(y_n) \prod_k p(l_n|p_k)p(p_k) \tag{7}$$

or in an additive setting

$$\underline{l}_n = \underline{y}_n + \sum_{k=1}^{K} h_{nk}\underline{p}_k \tag{8}$$

the random variables $\underline{y}_n \sim \mathcal{N}(\mu_n, \sigma^2_{e_n})$ and $\underline{p}_k \sim \mathcal{N}(0, \sigma^2_{p_k})$ being statistically independent. Variance propagation yields the full rank covariance matrix

$$\Sigma_{ll} = \text{Diag}([\sigma^2_{y_n}]) + \mathsf{H}\,\text{Diag}([\sigma^2_{p_k}])\mathsf{H}^\mathsf{T} \tag{9}$$

with $\mathsf{H} = [h_{nk}]$. The number of parameters for specifying this model is in the order of $O(NK)$, making the model specification efficient if $K \ll N$. The random variable $\underline{\boldsymbol{l}}$ can easily be described by the graphical model in Fig. 5.

The exchangeability of the functional (8) and the stochastical model, using Σ_{ll} from (9), has been discussed in the early days of aerial triangulation, see ACKERMANN (1965) and SCHILCHER (1980).

The graphical model makes the difference of modelling systematic errors, as done in bundle adjustment, transparent: Systematic errors usually are modelled with the second term $\sum_{k=1}^{K} h_{nk}\underline{p}_k$, then the causing effects are made explicit. But they also can be modelled using a fully correlated observation vector with covariance matrix Σ_{ll} in (9), see SCHILCHER (1980), then not making the causing effects explicit. Modelling systematic errors in the functional model or the stochastical model only is equivalent, if the expected values $\mathbb{E}(\underline{p}_k) = 0$ for the systematic effects are zero and constant over time. Obviously, modelling the systematic errors in the functional model not only requires less parameters to be specified compared to modelling them in the stochastical model, but also allows to model time dependent effects, see SCHROTH (1986) and to estimate their mean values and thus *learn* the parameters from data within a self-calibration.

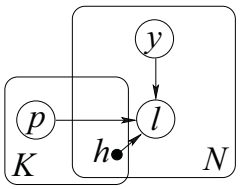

Fig. 5: Graphical model for describing the uncertainty of the N correlated random variables $\underline{l}_n$ using the primary error concept: The individual elements $\underline{l}_n$ depend on individual random variables $\underline{y}_n$ and common systematic effects $\underline{p}_k$ weighted by h_{nk}. The rounded boxes indicate that the internal structure, namely $\underline{l}$ depending on $\underline{y}$ and h on one hand and $\underline{p}$ and h on the other hand, are repeated N and K times respectively as made explicit in (8).

2.3.3 The Gauss-Markov model

The Gauss-Markov model is the work horse in statistical estimation. In its simplest form it could be written as $p(\boldsymbol{l}|\boldsymbol{x})$ only specifying the conditional probability for the observations given the parameters, thus being of type b) in Fig. 3. As soon as the observations $\boldsymbol{l}$ are available, we obtain the graphical model d), and using the Bayesian theorem we can derive the unknown parameters as $\text{argmax}_x p(\boldsymbol{x}|\boldsymbol{l}) = \text{argmax}_x p(\boldsymbol{l}|\boldsymbol{x})p(\boldsymbol{x})$ in case some prior information $p(\boldsymbol{x})$ about the parameters $\boldsymbol{x}$ is available. Self-calibrating bundle adjustment could be modelled a bit more expressivly, making the effect of the three parameter types, scene coordinates $\boldsymbol{k}$, orientation parameters $\boldsymbol{t}$, and additional parameters $\boldsymbol{p}$ onto the observed image coordinates $\boldsymbol{l}$ explicit, see Fig. 6.

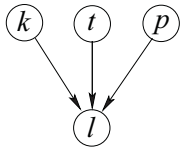

Fig. 6: Graphical model for bundle adjustment with coordinates $\boldsymbol{k}$, orientation parameters $\boldsymbol{t}$ and additional parameters $\boldsymbol{p}$ influencing the observations $\boldsymbol{l}_i$, e.g. the image coordinates. The joint probability is $p(\boldsymbol{l}, \boldsymbol{k}, \boldsymbol{t}, \boldsymbol{p}) = p(\boldsymbol{k})p(\boldsymbol{t})p(\boldsymbol{p}) \prod_i p(\boldsymbol{l}_i|\boldsymbol{k}, \boldsymbol{t}, \boldsymbol{p})$

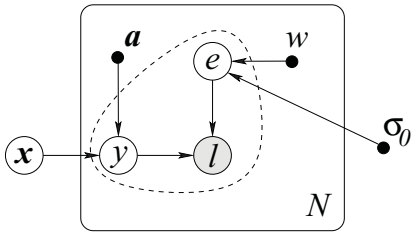

Fig. 7: Gauss-Markov model as Bayesian network. The black nodes represent given information, here the coefficients $\boldsymbol{a}$ of the (rows of the) design matrix, the weights w and the variance factor σ_0. The observations l result from a contamination of the unknown predicted observations, denoted by y, and the unknown measurement deviations e. The rounded box indicates the probabilistic model within the box is repeated N-times, indicating the independence of the N observed values l. In addition to the N fitted observations y and the measurement deviations e the parameter vector $\boldsymbol{x}$, which couples the fitted observations via the given coefficients $\boldsymbol{a}$, is unknown. Observe, this model contains the additive filtering model as core.

In case we want to model more details of the Gauss-Markov model, see Fig. 7, we can specify the dependency on the rows $\boldsymbol{a}_n$ of the design matrix, make the error free observations $y_n = \boldsymbol{a}_n^\mathsf{T}\boldsymbol{x}$ and the measurement deviations e_n explicit, and specify the individual weights w_n and the common factor σ_0 to obtain

$$\underline{l}_n = \boldsymbol{a}_n^\mathsf{T}\boldsymbol{x} + \underline{e}_n \qquad \sigma_n^2 = \sigma_0^2/w_n \tag{10}$$

together with the prior information

$$\underline{\boldsymbol{x}} \sim \mathcal{N}(\boldsymbol{x}_0, \Sigma_{x_0x_0}) \tag{11}$$

e.g. representing the coordinates of the control points together with their uncertainty, assuming very large covariances for the new unknown scene points. The joint probability is given in the factorized form

$$p(\boldsymbol{l}, \boldsymbol{x}) = p(\boldsymbol{x}) \prod_n p(l_n|\boldsymbol{x}) \,. \tag{12}$$

Observe, the Gauss-Markov model contains the filtering model from Fig. 4 as central part within the area with dashed boundary in Fig. 7, here in the form $l = y + e$.

We will not discuss general Bayesian nets here. Finding the optimal parameters in Bayesian nets with a tree structure, such as all examples, is linear in the number of nodes, if the number of variables per node is fixed. This high efficiency is exploited e.g. in Kalman filtering.

2.4 *Markov Random Fields*

Markov random fields are graphical models with undirected edges. Undirected graphs appear in image processing, image analysis, point cloud processing, but also in geodetic and photogrammetric networks.

In image processing the graph is induced by the regular structure of the pixels, yielding a regular pattern, see Fig. 8, lower left. In image analysis one often starts with partitioning the image into regions, the region adjacency

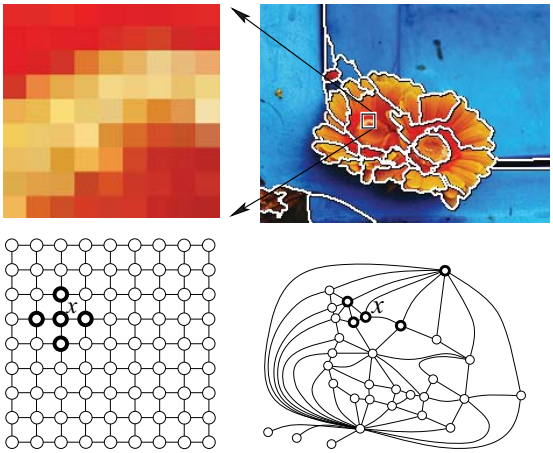

Fig. 8: Graphs of photogrammetric models as basis for Markov random fields. *Left:* pixel grid, neighbouring or adjacent pixels are assumed to have the same colour or the same class label. *Right:* region adjacency graph, neighbouring regions are assumed to have the same class label (image and segmentation from SCHINDLER & FÖRSTNER (2013)).

graph then may be the building block for image interpretation, which is equivalent to assigning a class name to each region, see Fig. 8, lower right. In point cloud processing the graph may result from a triangulation or a segmentation. Bundle adjustment, see Fig. 9 links the observed image coordinates $\boldsymbol{x}_{ij}$, not to be confused with the unknown parameters, with scene coordinates $\boldsymbol{k}_i$ and orientation parameters $\boldsymbol{t}_j$ via the collinearity constraints $0 = \text{coll}_{ij}(\boldsymbol{x}_{ij}, \boldsymbol{k}_i, \boldsymbol{t}_j)$.

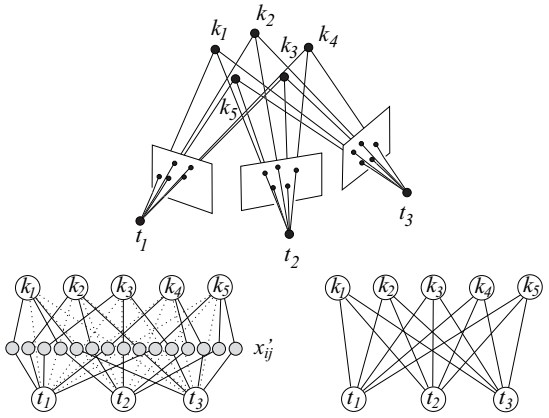

Fig. 9: Graphs of photogrammetric models as basis for Markov random fields. Bundle adjustment. *Lower left*: with 3-cliques (cliques with three nodes) linking the observations x_{ij} with the coordinates and the transformation parameters (some of the connections are shown dotted to keep the figure readable). *Lower right:* with 2-cliques (cliques with two nodes) linking coordinates and transformation parameters, observations as given values are eliminated, see DELLAERT & KAESS (2006).

A Markov random field having nodes x_i is characterized by the following property: The probability for a node $p(x_i|\mathcal{X} \setminus x_i)$ given the values of all other nodes is identical to the conditional probability $p(x_i|\mathcal{N}_i)$ given its neighbours:

$$p(x_i|\mathcal{X} \setminus x_i) = p(x_i|\mathcal{N}_i)\,. \qquad (13)$$

This is a generalization of the Markov property of time series where the current state x_t only depends on the previous thus neighbouring one x_{t-1}, explaining the name for the type of graphical model.

The joint probability $p(x_1, ..., x_N)$ of all variables x_n in a Markov random field can be written as the product of factors $\psi_i(\mathcal{X}_i)$ depending on the cliques in the corresponding graph, the clique $\mathcal{C}_i$ being a fully connected subgraph, i.e. a subgraph, where all nodes are connected, inducing the set $\mathcal{C}_i = \{x_{i_1}, ..., x_{i_{N_i}}\}$:

$$p(x_1, ..., x_N) = \frac{1}{Z} \prod_{i \in C} \psi_i(\mathcal{C}_i)\,. \qquad (14)$$

The functions $\psi_i(\mathcal{C}_i)$ are called potential functions and are assumed to be positive for all values the variables in the clique may have. They conceptually are *no probabilities*.

A probability of the structure (14) is called a Gibbs-distribution. The structure of this distribution is equivalent to the Markov property in (13). This follows from a theorem by HAMMERSLEY & CLIFFORD (1971), see also BUSCH (1992). In Fig. 8 the lower left network for the image grid shows only two-cliques, the lower right network of the region adjacency graph shows two- and three-cliques and the network for the adjustment in Fig. 9 shows three-cliques or two-cliques, depending on whether the observational nodes are included or excluded in the model.

2.4.1 Types of Markov random fields

There are different classes of Markov random fields (MRF), depending on the way observations are handled. For simplicity we refer to the grid type graph with maximal two-cliques, see Fig. 10.

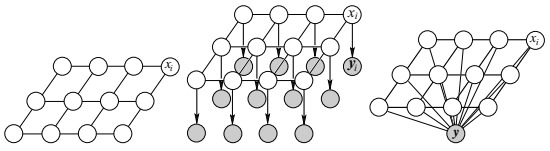

Fig. 10: Types of MRF's. *Left:* Markov random field as prior. *Middle:* hidden Markov random field, containing directed edges to the observations y_i. *Right:* Conditional random field.

A Markov random field may be used for modelling the prior distribution of the image:

$$p(\boldsymbol{x}) = \frac{1}{Z} \prod_{(i,j) \in \mathcal{E}} \psi_{ij}(x_i, x_j) \qquad (15)$$

where only the potential functions $\psi_{ij}(x_i, x_j)$ need to be specified such that likely configurations of pairs (x_i, x_j) obtain large values $\psi_{ij}(x_i, x_j) > 0$. The key publication by GEMAN & GEMAN (1984) of the school of Grenander (GRENANDER 1976), addressed the problem of image restoration as a global optimization problem solved by stochastic relaxation in a Markov Chain Monte Carlo scheme.

A hidden Markov random field makes explicit the fact, that the class labels for the pixels are unknown. The naming disregards the fact that the edges to the observed nodes are directed. The observations y_i depend on the corresponding class labels x_i. Thus, the total probability $p(\boldsymbol{x}, \boldsymbol{y})$ of the network is

$$p(\boldsymbol{x})p(\boldsymbol{y}|\boldsymbol{x}) = \frac{1}{Z}\prod_i p(y_i|x_i) \prod_{(ij)\in\mathcal{E}} \psi_{ij}(x_i, x_j) \quad (16)$$

where the prior $p(\boldsymbol{x})$ is the simple Markov random field (15). The likelihood $p(\boldsymbol{y}|\boldsymbol{x})$ can be factorized into a product of factors $\psi_i(x_i, y_i) := p_i(y_i|x_i)$ for given observations only depending on individual nodes.

Often one wants to use observations y_i which are taken from a region around the i-th pixel. Then they are not any more independent. This is the reason why one uses

Conditional random fields proposed by LAFFERTY et al. (2001) in the area of language processing and introduced to image analysis problems by KUMAR & HEBERT (2006) under the label *discriminative random fields*, though already e.g. GIMEL'FARB (1996) used models of this structure for texture analysis. The main idea is the following: one directly models the posterior probability by conditioning the probabilistic relations in the cliques, here between neighbouring pixels, on the given observations

$$p(\boldsymbol{x}|\boldsymbol{y}) = \frac{1}{Z}\prod_{i\in C_1} \psi_i(x_i, \boldsymbol{y}) \prod_{(ij)\in\mathcal{C}_2} \psi(x_i, x_j, \boldsymbol{y}) \quad (17)$$

where the one-cliques $\mathcal{C}_1$ contain all nodes, and the two-cliques $\mathcal{C}_2$ all edges.

In all cases it is useful to replace the product of the probabilities and potentials by a sum of the negative logarithms. The negative logarithm of a probability can be interpreted as the (self-) information $I(x) = -\log p(x)$, being the relative surprise when observing the sample (SHANNON & WEAVER 1949). Correspondingly the negative logarithm $E(x) = -\log \psi(x)$ of the potentials is called the energy, as the concept has been developed in physics. Thus, the complete models can be written as sums.

When neglecting the given observed, fixed values $\boldsymbol{y}$, for the conditional random field model (17) we obtain

$$I(\boldsymbol{x}|\boldsymbol{y}) = \log Z + \sum_{i\in C_1} E_i(x_i) + \sum_{(ij)\in\mathcal{C}_2} E_{ij}(x_i, x_j)\,. \quad (18)$$

2.4.2 Tasks

Having fixed the structure of the model several tasks need to be addressed.

Inference. The first task is inference: In case the model is fully specified, one is interested in an estimate $\widehat{\boldsymbol{x}}$ for the optimal set of values for the unknowns $\{x_i\}$, e.g. the one which maximizes the probability $p(\boldsymbol{x}|\boldsymbol{y})$, leading to the maximum a posteriori (MAP) estimate, or, equivalently, the one which minimizes $I(\boldsymbol{x}|\boldsymbol{y})$. If the x_i represent class labels in an image, this will lead to an optimal labelling of all pixels or regions of the image. If the x_i represent the continuous values of the intensities of the pixels, one might obtain an optimal restoration of the image.

The problem in general is intractable, i.e. the computing time increases exponentially with the number of nodes, which is plausible when regarding the large search space, see the discussion of the number of images above. Only in very special cases finding the optimum is tractable: (a) in case the variables are continuous and follow a Gaussian distribution, the optimal solution can be achieved by solving a large linear equation system, (b) in case the variables are binary and the potential functions obey certain conditions, the problem can be

mapped to a network algorithm, namely finding the maximum flow or the minimal cut in a network (BOYKOV & KOLMOGOROV 2004). In all other cases only suboptimal algorithms exist (BISHOP 2006).

None of the algorithms yields the probability $p(\boldsymbol{x}|\boldsymbol{y})$ as this would require the determination of the normalization constant Z, which is $Z = \sum_{\boldsymbol{x}} p(\boldsymbol{x}|\boldsymbol{y})$, the sum over all possible images.

One simple algorithm uses the Markov property and iteratively determines the locally best estimate given all neighbours:

$$\widehat{x}_i^{(\nu+1)} = \text{argmax}_{x_i} p\left(x_i \mid \mathcal{N}_i^{(\nu)}\right) \tag{19}$$

where (ν) indicates the iteration number. The algorithm is known as *iterative conditional modes* (ICM). It often yields good results, but may get stuck in a local minimum of $I(\boldsymbol{x}|\boldsymbol{y})$. We will relate this algorithm to one used in adjustment theory.

As graphical models factor the joint probability into small factors, see (1) and its specialization for Bayesian nets and Markov random fields, see (16), they can be represented as so-called factor graphs, which allow for a homogeneous algorithmic handling of Bayesian nets and Markov random fields (KSCHISCHANG et al. 2001).

Learning. The second task is learning. It results from the difficulty in specifying the potentials in real world problems completely. Take an example: As we are free to choose the potential functions we could use posterior probabilities $p(x_i|\boldsymbol{y}_i)$ and $p(x_i, x_j|\boldsymbol{y}_{ij})$ of classifiers for labels x_i or label pairs (x_i, x_j). Both depend on the corresponding observed features $\boldsymbol{y}_i$ and $\boldsymbol{y}_{ij}$, and possibly global parameters $\boldsymbol{y}_g$, collected in the vector $\boldsymbol{y}$. Usually they are parametrized, e.g. in case one uses a maximum likelihood classifier based on a Gaussian distribution. Then the parameter vector $\boldsymbol{u}$ would contain the mean vectors and covariance matrices of the classifier to be learnt. Thus, we would obtain

$$\begin{aligned} &\psi_i(x_i|\boldsymbol{u}) \\ &= p(x_i|\boldsymbol{y}_i, \boldsymbol{u}_1)\psi_{ij}(x_i, x_j|\boldsymbol{y}_{ij}, \boldsymbol{u}_2) \end{aligned} \tag{20}$$

where the parameters $\boldsymbol{u} = (\boldsymbol{u}_1, \boldsymbol{u}_2)$ need to be learned from the data. The complete model, now written using negative logarithms then is

$$\begin{aligned} &I(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{u}) = \log Z(\boldsymbol{u}) + \\ &+ \sum_{i \in C_1} E_i(x_i, \boldsymbol{u}_1) + \sum_{(ij) \in \mathcal{C}_2} E_{ij}(x_i, x_j, \boldsymbol{u}_2) \end{aligned} \tag{21}$$

where the normalization factor $1/Z(\boldsymbol{u})$ also depends on the unknown parameters. Given a large enough set of training data $(x_i, \boldsymbol{y}_i)$ and $(x_i, x_j, \boldsymbol{y}_{ij})$ the task is to learn, i.e. to estimate optimal parameters for $\boldsymbol{u}$. Again, the problem of estimating optimal parameters is intractable in general and only suboptimal solutions are known.

The following example wants to demonstrate the possibility to map classical geodetic networks to Markov random fields

2.5 *A Four-Node Network*

In order to demonstrate the flexibility of graphical models, especially of conditional random fields, we use a four-node graph with one edge missing as an example. We do not show the observational nodes, which would be linked to all four nodes, see Fig. 11.

2.5.1 A levelling network

The graph with four nodes representing the 1-cliques $\mathcal{C}_1 = \{1, 2, 3, 4\}$ and five edges representing the 2-cliques $\mathcal{C}_2 = \{(1, 2), (2, 3), (3, 4), (1, 4), (2, 4)\}$, has two 3-cliques $\mathcal{C}_3 = \{(1, 2, 4), (2, 3, 4)\}$ as there are only two fully connected triangles. Interpreting the graph as a levelling network restricts the modelling to 1- and 2-cliques, as we only have measured heights $l_i, i = 1, 3$ of the two control points and the five height differences $l_{ij}, (i, j) \in \mathcal{C}_2$.

We now model the joint probability with *maximal cliques*, thus by 2-cliques only. This is achieved by taking the potentials for the observed control points into on the potential of one of the corresponding the 2-cliques. The joint probability therefore is

$$\begin{aligned} p(\boldsymbol{x}) &= p(x_1, x_2, x_3, x_4) \\ &\propto \prod_{(i,j) \in \mathcal{C}_2} \psi_{ij}(x_i, x_j) \\ &\propto \psi_{12}(x_1, x_2) \bullet \ldots \bullet \psi_{34}(x_3, x_4)\,. \end{aligned}$$

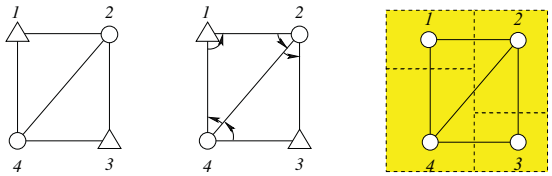

Fig. 11: *Left:* Leveling network as a conditional random field. The triangles indicate reference points with given heights h_1 and h_3, in principle being uncertain. The two unknown heights h_2 and h_4 are indicated by white circles. The edges in the graph indicate probabilistic relations between the two nodes in concern, which practically express the uncertainty of the five observed height differences. No observations relate three heights simultaneously, i.e. only two-cliques in the graph give rise to potentials. *Middle:* Angular network as a conditional random field. The triangles now represent control points in 2D with their uncertain Cartesian coordinates $\boldsymbol{x}_1$ and $\boldsymbol{x}_3$. The coordinates $\boldsymbol{x}_2$ and $\boldsymbol{x}_4$ are unknown. The five edges in the graph again represent probabilistic relations between the corresponding nodes. Each angle measurement is indicated by arrows and depends on three points. Only two three-cliques in the graph give rise to potentials. *Right:* Classification of four parcels as a conditional random field. The nodes represent the unknown classes of the four fields. There may be prior probabilities on the classes, known from the type of geographic region. The edges again probabilistically constrain the corresponding classes, stating neighbouring fields to more likely belong to the same class than to different classes, which may be made dependent on observations, e.g. of the appearance of the field boundaries.

In order to choose adequate potential functions we assume the heights and the height differences to be normally distributed

$$\begin{aligned} \underline{l}_i &\sim \mathcal{N}(\mu_{x_i}, w_{x_i}^{-1}), i \in \mathcal{C}_1 \\ \underline{l}_{ij} &\sim \mathcal{N}(\mu_{x_j} - \mu_{x_i}, w_{ij}^{-1}), \quad (i,j) \in \mathcal{C}_2 \end{aligned} \tag{22}$$

where we used the weights $w = 1/\sigma^2$. Therefore, we choose the following potential functions, being able to neglect the constant factors of the normal distribution

$$\begin{aligned} \psi_{12} &= \mathrm{e}^{-\frac{1}{2}(x_2-x_1-l_{12})^2 w_{12}} \mathrm{e}^{-\frac{1}{2}(x_1-l_1)^2 w_1} \\ \psi_{14} &= \mathrm{e}^{-\frac{1}{2}(x_4-x_1-l_{14})^2 w_{14}} \\ \psi_{23} &= \mathrm{e}^{-\frac{1}{2}(x_3-x_2-l_{23})^2 w_{23}} \\ \psi_{24} &= \mathrm{e}^{-\frac{1}{2}(x_4-x_2-l_{24})^2 w_{24}} \\ \psi_{34} &= \mathrm{e}^{-\frac{1}{2}(x_4-x_3-l_{34})^2 w_{34}} \mathrm{e}^{-\frac{1}{2}(x_3-l_3)^2 w_1} . \end{aligned} \tag{23}$$

Observe, we only used 2-cliques and integrated the prior information about the control points into one of the neighbouring 2-cliques. The potential functions essentially depend on the weighted squares of the differences v_i and v_{ij} between the unknown parameters and the measurements. Taking negative logarithms we therefore obtain the – not really surprising – result: We need to find the minimum of

$$I(\boldsymbol{x}|\boldsymbol{l}) = \sum_{(i,j)\in\mathcal{C}_2} w_{ij} v_{ij}^2(x_i, x_j) \tag{24}$$

here with $v^2 := -2\log\psi$, thus v_{12}^2 and v_{34}^2 containing the squared residuals at the control points. *The leveling network is a special Markov random field, namely a Gaussian random field.*

We now apply the iterative conditional mode algorithm to this problem. The normal equation system, resulting from setting $\partial I/\partial \boldsymbol{x} = \mathbf{0}$ and assuming weights $w_i = w_{ij} = 1$ is

$$\begin{bmatrix} \mathbf{3} & -1 & \mathbf{0} & -1 \\ -1 & 3 & -1 & -1 \\ \mathbf{0} & -1 & \mathbf{3} & -1 \\ -1 & -1 & -1 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \\ = \begin{bmatrix} -l_{12} - l_{14} + l_1 \\ l_{12} - l_{23} - l_{24} \\ l_{23} - l_{34} + l_3 \\ l_{34} + l_{14} + l_{24} \end{bmatrix} = \begin{bmatrix} h_1 \\ h_2 \\ h_3 \\ h_4 \end{bmatrix} . \tag{25}$$

The best estimate for the individual unknowns, given the others, therefore are

$$\begin{aligned} x_1^{(\nu+1)} &= \frac{1}{3}(h_1 + x_2^{(\nu)} + x_4^{(\nu)}) \\ x_2^{(\nu+1)} &= \frac{1}{3}(h_2 + x_1^{(\nu+1)} + x_3^{(\nu)} + x_4^{(\nu)}) \\ x_3^{(\nu+1)} &= \frac{1}{3}(h_3 + x_2^{(\nu+1)} + x_4^{(\nu)}) \\ x_4^{(\nu+1)} &= \frac{1}{3}(h_4 + x_1^{(\nu+1)} + x_3^{(\nu+1)} + x_2^{(\nu+1)}) \end{aligned}$$

again with the iteration index (ν). This method of solving the normal equations iteratively is known as Gauss-Seidel method. It is guaranteed to converge in this case, as the normal equation matrix is symmetric and positive definite. This criterion of course is difficult to generalize for non-Gaussian situations. This procedure is equivalent to the so-called method of

iterative conditional modes for iteratively finding the optimal solution in a general Markov-random field using (13) in the form (BUSCH 1992)

$$x_i^{(\nu+1)} := \text{argmax}_{x_i} p(x_i \mid \mathcal{N}_i^{(\nu)}(x_i)) \,. \quad (26)$$

A final remark refers to the sparsity of the model. In spite of the small size of the network the graph is not fully connected, as edge $(1,3)$ is missing. This is reflected in the zero-entry $N_{13} = 0$ of the normal equation matrix. The zero at the $(1,3)$ position indicates that given the other variables x_2 and x_4 the variables x_1 and x_3 are independent: formally $p(x_1, x_3|x_2, x_4) = p(x_1|x_2, x_4)p(x_3|x_2, x_4)$, as then the resulting normal equation system for x_1 and x_3 is diagonal. This is indicated by the boldtype numbers in the normal equation matrix in (25). The conditional independence results in a sparse graphical model. In adjustment theory it regularly is exploited for increasing the efficiency of computation of large networks. The sparseness of the normal equation matrix however is an indication for the sparsity of the complete model resulting from the type of measuring design, which therefore in a natural way leads to a sparse graphical model.

2.5.2 An angular network

We now use the same graph to represent a 2D network with measured angles. The nodes then represent random 2-vectors $\boldsymbol{x}_i$. The observations l_{ijk} are the angles $l_{ijk} := \alpha_{ijk} = \phi_{jk} - \phi_{ji}$ and depend on the coordinates of the points. Thus, we will need the two 3-cliques to represent the probability of the complete network:

$$p(\boldsymbol{x}|\boldsymbol{l}) = \frac{1}{Z}\psi_{124}(\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_4)\psi_{234}(\boldsymbol{x}_2, \boldsymbol{x}_3, \boldsymbol{x}_4) \,. \quad (27)$$

We assume the control point coordinates to be Gaussian distributed

$$\underline{\boldsymbol{l}}_i \sim \mathcal{N}(\boldsymbol{\mu}_{x_i}, w_i^{-1}\boldsymbol{I}_2) \,, i = 1, 3 \,. \quad (28)$$

Again we assume the angular measurements to be normally distributed

$$\underline{l}_{ijk} \sim \mathcal{N}(\alpha_{ijk}(\boldsymbol{x}_i, \boldsymbol{x}_j, \boldsymbol{x}_k), w_{ijk}^{-1}) \quad (29)$$

with $(i, j, k) \in \mathcal{A} = \{$(4,1,2), (1,2,4), (2,4,1), (3,4,2), (4,2,3)$\}$. This yields the potentials

$$\psi_{124}(\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_4) = \quad (30)$$
$$\mathrm{e}^{-\frac{1}{2}\left(|\alpha_{421}-l_{421}|^2 w_{421}+|\alpha_{142}-l_{142}|^2 w_{142}\right)}$$
$$\bullet\mathrm{e}^{-\frac{1}{2}\left(|\alpha_{214}-d_{214}|^2 w_{214}+|x_1-l_1|^2 w_1\right)}$$

and

$$\psi_{234}(\boldsymbol{x}_2, \boldsymbol{x}_3, \boldsymbol{x}_4) = \quad (31)$$
$$\mathrm{e}^{-\frac{1}{2}\left(|\alpha_{243}-l_{243}|^2 w_{243}+|\alpha_{324}-l_{324}|^2 w_{324}\right)}$$
$$\bullet\mathrm{e}^{-\frac{1}{2}\left(|x_3-l_3|^2 w_3\right)} \,.$$

2.5.3 A classification network

We finally use the graph for modelling a classification task. Let four agricultural fields be arranged as shown in the Fig. 11 right. The graph then represents the region adjacency.

One usually takes some spectral or texture features $\boldsymbol{y}_i$ in each region from the underlying image (not shown) as observations, and performs a Bayesian classification, i.e. labelling of the regions. This is achieved by maximizing the posterior probability $p(x_i|\boldsymbol{y}_i) \propto p(\boldsymbol{y}_i|x_i)p(x_i)$ using the likelihood $L(x_i) := p(\boldsymbol{y}_i|x_i)$, which is to be learnt from training data and some prior $p(x_i)$ on the occurrence of the different classes $x_i \in \{1, 2, ...\}$. It corresponds to four independent Bayesian nets of the type Fig. 3 d. The joint prior of all four nodes just is $p(\boldsymbol{x}) = \prod_{i=1}^{4} p(x_i)$, as the class membership or the labelling of the regions is assumed to be independent.

In case one can assume neighbouring regions often belong to the same class, a simple model for the prior would therefore be to assume neighbouring regions are likely belonging to the same class, and unlikely belonging to different classes. Assuming the probability for the complete network for both cases should differ by a factor 10, this can be expressed by the potential function for each edge

$$\psi(x_i, x_j) = \begin{cases} 1, \text{ if } x_i = x_j \\ 0.1, \text{ if } x_i \neq x_j \end{cases} \,. \quad (32)$$

The prior for all nodes then is

$$p(\boldsymbol{x}) = \frac{1}{Z}\prod_{(ij)\in\mathcal{C}_2} \psi(x_i, x_j) \quad (33)$$

which leads to the complete probability

$$p(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{Z} \prod_{i \in \mathcal{C}_1} p(x_i|\boldsymbol{y}) \prod_{(ij) \in \mathcal{C}_2} \psi(x_i, x_j)\,. \tag{34}$$

This is the so-called Potts model for the prior in classification using a Markov random field. As discussed above, finding an optimal labelling is intractable. Examples with a small network are given by KORČ (2012).

3 Outlook and Conclusion

Graphical models are a powerful tool for communicating between users of a statistical model and experts in statistics. The paper showed various examples for models used in geodesy and photogrammetry where graphical models give insight into the internal structure of the models, especially making fixed, observed and unknown parameters explicit and showing the independence assumptions made.

For a graphical model, where the relationships are linear and the distributions are Gaussian, reasoning, especially optimization, leads to a linear equation systems guaranteeing a unique solution. For statistical optimization problems resulting from a graphical model in general no unique method exists. This does not only hold for parameter estimation problems, e.g. in the presence of outliers, but – with only few exceptions – for all classification problems due to the discrete nature of the search of the parameter space.

The paper shows various relations between modelling and estimation in geodetic and photogrammetric networks. The parsimony of the used models results from the special structure of most networks. Interpreting such networks as Markov random fields allows one to see the more general structure of the problems at hand. Examples are the modelling of cycle slips in GPS observations, where discrete and continuous variables occur simultaneously, or the change of the type of distribution, e.g. from Gaussian to Laplace, or the type of prior knowledge. These abilities not only build a bridge between models in geometry and in classification, but also between photogrammetry and remote sensing.

Acknowledgements

I thank the reviewers for their very careful work.

References

ACKERMANN, F., 1965: Fehlertheoretische Untersuchungen über die Genauigkeit photogrammetrischer Streifentriangulationen. – Deutsche Geodätische Kommission bei der Bayerischen Akademie der Wissenschaften **C 87**, München.

BILMES, J., 2000: Introduction to Graphical Models. http://www.ee.washington.edu/class/596/patrec/ (1.3.2013).

BISHOP, C., 2006: Pattern Recognition and Machine Learning. – Springer-Verlag, NY, USA.

BOYKOV, Y. & KOLMOGOROV, V., 2004: An Experimental Comparison of Min-Cut/Max-Flow Algorithms for Energy Minimization in Vision. – IEEE Transactions on Pattern Analysis and Machine Intelligence **26** (9): 1124–1137.

BUSCH, A., 1992: Bayes-Statistik und Markoff-Felder für die Restaurierung digitaler Bilder. – Deutsche Geodätische Kommission bei der Bayerischen Akademie der Wissenschaften **C 396**, München.

DELLAERT, F. & KAESS, M., 2006: Square Root SAM: Simultaneous localization and mapping via square root information smoothing. – International Journal of Robotics Reasearch **25**: 2006.

GEMAN, S. & GEMAN, D., 1984: Stochastic Relaxation, Gibbs distributions and the Bayesian restoration of images. – IEEE T-PAMI **6** (6): 721–741.

GIMEL'FARB, G.L., 1996: Texture Modeling by Multiple Pairwise Pixel Interactions. – IEEE T-PAMI **18** (11): 1110–1114.

GRENANDER, U., 1976: Lectures in Pattern Theory. – Volumes I-III of Applied Mathematical Sciences, Springer Verlag, New York, USA.

HAMMERSLEY, J.M. & CLIFFORD, P., 1971: Markov fields on finite graphs and lattices. – unpublished, Berkeley, CA, USA. http://www.statslab.cam.ac.uk/˜grg/books/hammfest/hamm-cliff.pdf (6.6.2013).

KORČ, F., 2012: Tractable Learning for a Class of Global Discriminative Models for Context Sensitive Image Interpretation. – PhD thesis, Department of Photogrammetry, University of Bonn.

KRAUS, K., 1972: Interpolation nach kleinsten Quadraten in der Photogrammetrie. – BuL **1**: 7–12.

KSCHISCHANG, F., FREY, B. & LOELIGER, H.-A., 2001: Factor graphs and the sum-product algorithm. – Information Theory, IEEE Transactions on **47** (2): 498–519.

KUMAR, S. & HEBERT, M., 2006: Discriminative Random Fields. – International Journal of Computer Vision **68** (2): 179–201.

LAFFERTY, J.D., MCCALLUM, A. & PEREIRA, F.C.N., 2001: Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. – **Eighteenth** International Conference on Machine Learning, ICML '01: 282–289, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

LI, S.Z., 2000: Markov random field modeling in computer vision. – Springer-Verlag, London, UK.

MORITZ, H., 1978: Least squares collocation. – Revision of Geophysics and Space Physics **16** (3): 421–430.

MURPHY, K., 1998: A Brief Introduction to Graphical Models and Bayesian Networks. http://www.cs.ubc.ca/~murphyk/Bayes/bnintro.html (1.3.2013).

PEARL, J., 1988a: Embracing causality in default reasoning. – Artificial Intelligence **35** (2): 259–271.

PEARL, J., 1988b: Probabilistic Reasoning in Intelligent Systems. – Morgan Kaufmann Publishers.

PELZER, H., 1974: Zur Behandlung singulärer Ausgleichungsaufgaben. – Zeitschrift für Vermessungswesen **5**: 181–194, 479–488.

RASMUSSEN, C.E. & WILLIAMS, C.K.I., 2005: Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning). – The MIT Press, Cambridge, MA, USA.

SCHILCHER, M., 1980: Empirisch-statistische Untersuchungen zur Genauigkeitsstruktur des photogrammetrischen Luftbildes. – Deutsche Geodätische Kommission bei der Bayerischen Akademie der Wissenschaften **C 262**, München.

SCHINDLER, F. & FÖRSTNER, W., 2013: Marching Front Graph Partitioning in Geometry and Image Processing. – PFG – Photogrammetrie, Fernerkundung, Geoinformation.

SCHROTH, R., 1986: Ein erweitertes mathematisches Modell der Aerotriangulation zur hochgenauen Punktbestimmung. – Deutsche Geodätische Kommission bei der Bayerischen Akademie der Wissenschaften **C 316**, München.

SHANNON, C.E. & WEAVER, W., 1949: The Mathematical Theory of Communication. – The University of Illinois Press. Urbana, IL, USA.

WINKLER, G., 2006: Image Analysis, Random Fields and Markov Chain Monte Carlo Methods: A Mathematical Introduction (Stochastic Modelling and Applied Probability). – Springer-Verlag New York, Secaucus, NJ, USA.

Address of the Authors:

Prof. Dr.-Ing. Dr. h. c. mult. WOLFGANG FÖRSTNER, Josef-Schell-Str. 34, 53121 Bonn, e-mail: wf@ipb.uni-bonn.de

Manuskript eingereicht: März 2013
Angenommen: Mai 2013