Recursive Cavity Modeling for Estimation of Gaussian MRFs*

Stochastic Systems Group

Jason K. Johnson

October 9, 2002

*/mit/jasonj/Public/SSG-OCT9-02

Overview

- Background
  - Graphical Models (MRFs)
  - Exponential Families
  - Gaussian MRFs
  - Information Geometry and Projections
- Model-Thinning Projections
  - Model Selection by greedy edge-removal procedure.
  - Parameters optimized by Iterative Scaling.
- Recursive Cavity Modeling
  - Nested Dissection
  - Cavity Modeling
  - Blanket Modeling
  - Examples

Graphical Models*

Undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ based upon vertices $\mathcal{V}$ with $\mathcal{E}$ (unordered pairs of vertices).

Random variables $\mathbf{x} = (x_i, i \in \mathcal{V})$ are said to be *Markov* w.r.t $\mathcal{G}$ when

$$p(x_A, x_B | x_S) = p(x_A | x_S) p(x_B | x_S)$$

for all $A, B, S \subset \mathcal{V}$ where S seperates A from B.

Hammersley-Clifford, 71.† $\mathbf{x}$ is Markov w.r.t. $\mathcal{G}$ if and only if $p(x)$ factors according to $\mathcal{G}$ as

$$p(x) = \frac{1}{Z(\psi)} \prod_{c \in \mathcal{C}} \psi_c(x_c)$$

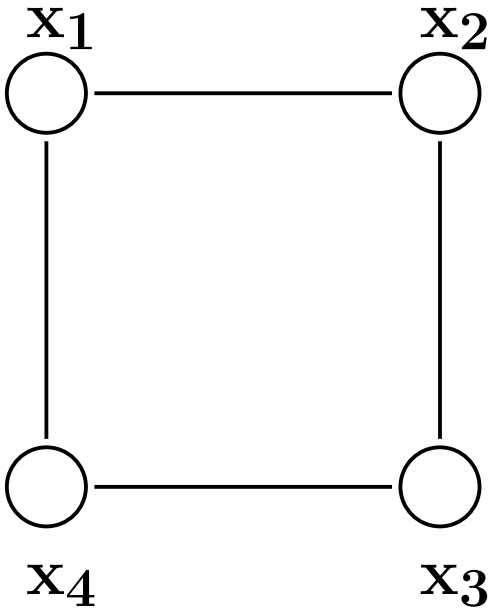
with positive *potential functions* ψ and $Z(\psi)$ is normalization constant.

Markov structure of random process $\mathbf{x}$ allows for compact specification of $p(x)$ as graphical models.

*Lauritzen, 96; Jordan, 99.

†Grimmett, 73.

Example MRF



Graph Factorization

$$\begin{aligned} \boldsymbol{p}(\boldsymbol{x}) \;\; \propto \;\; & \boldsymbol{\psi_1}(\boldsymbol{x_1})\boldsymbol{\psi_2}(\boldsymbol{x_2})\boldsymbol{\psi_3}(\boldsymbol{x_3})\boldsymbol{\psi_4}(\boldsymbol{x_4}) \\ & \boldsymbol{\psi_{1,2}}(\boldsymbol{x_1},\boldsymbol{x_2})\boldsymbol{\psi_{2,3}}(\boldsymbol{x_2},\boldsymbol{x_3}) \\ & \boldsymbol{\psi_{3,4}}(\boldsymbol{x_3},\boldsymbol{x_4})\boldsymbol{\psi_{4,1}}(\boldsymbol{x_4},\boldsymbol{x_1}) \end{aligned}$$

Conditional Independence

$$\boldsymbol{p}(\boldsymbol{x_{1,3}}|\boldsymbol{x_{2,4}}) = \boldsymbol{p}(\boldsymbol{x_1}|\boldsymbol{x_{2,4}})\boldsymbol{p}(\boldsymbol{x_3}|\boldsymbol{x_{2,4}})$$

$$\boldsymbol{p}(\boldsymbol{x_{2,4}}|\boldsymbol{x_{1,3}}) = \boldsymbol{p}(\boldsymbol{x_2}|\boldsymbol{x_{1,3}})\boldsymbol{p}(\boldsymbol{x_4}|\boldsymbol{x_{1,3}})$$

Exponential Families*

Specified by a *base measure* $q(x) > 0$ and a set of *sufficient statistics* $t(x)$ both defined over some specified state-space X. We take $X = R^n$ so that model is specified by pdf of the form

$$f(x;\theta) = q(x)\exp\{\theta \cdot t(x) - \varphi(\theta)\}$$

where the *cumulant function* $\varphi(\theta)$ is the normalization constant

$$\varphi(\theta) = \log \int q(x)\exp\{\theta \cdot t(x)\}dx$$

Only consider *admissable* parameters Θ s.t. pdf is normalizable $\varphi(\theta) < \infty$. The family is *regular* if Θ has non-empty interior. The statistics are *minimal* if the $t(x)$ are linearly-independent. Then, dual parameterization provided by *moment coordinates* $\eta = E_\theta\{t(\mathrm{x})\}$ over the set of achievable moments $\eta(\Theta)$.

*Chentsov, 66; Barndorff-Nielsen, 78.

Gaussian Markov Random Fields

Consider Gaussian process $\mathbf{x} \sim \mathcal{N}(\mu, \Sigma)$ with mean vector $\mu = E\{\mathbf{x}\}$ and covariance matrix $\Sigma = E\{\mathbf{x}\mathbf{x}'\} - \mu\mu'$.

Information Filter Form. Say that $\mathbf{x} \sim \mathcal{N}^{-1}(h, J)$ if

$$\begin{aligned} h &= \Sigma^{-1}\mu \\ J &= \Sigma^{-1} \end{aligned}$$

s.t. density function is parameterized as

$$p(x) = \exp\{-\frac{1}{2}x'Jx + h'x - \varphi(h, J)\}$$

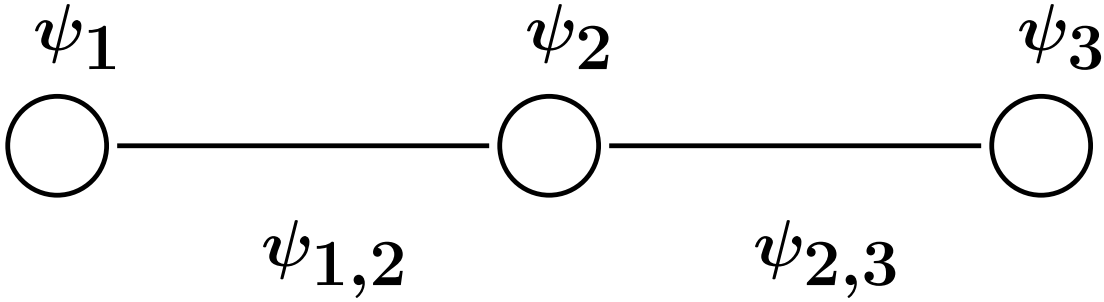
where

$$\varphi(h, J) = \frac{1}{2}\{h'J^{-1}h - \log|J| + n\log 2\pi\}.$$

This is an exponential family model with

$$\begin{aligned} \theta &= (h, -J/2) \\ t(x) &= (x, xx') \\ \eta &= (\mu, \Sigma + \mu\mu') \\ \varphi(\theta) &= \varphi(h, J) \end{aligned}$$

Example GMRF



$$p(x) \propto \psi_1(x_1)\psi_2(x_2)\psi_3(x_3)\psi_{1,2}(x_1,x_2)\psi_{2,3}(x_2,x_3)$$

$$\psi_1(x_1) = \exp\{-\frac{1}{2}x_1' J_{1,1} x_1 + h_1' x_1\}$$

$$\psi_2(x_2) = \exp\{-\frac{1}{2}x_2' J_{2,2} x_2 + h_2' x_2\}$$

$$\psi_3(x_3) = \exp\{-\frac{1}{2}x_3' J_{3,3} x_3 + h_3' x_3\}$$

$$\psi_{1,2}(x_1,x_2) = \exp\{-x_1' J_{1,2} x_2\}$$

$$\psi_{2,3}(x_2,x_3) = \exp\{-x_2' J_{2,3} x_3\}$$

$$h = \begin{pmatrix} h_1 \\ h_2 \\ h_3 \end{pmatrix}, \quad J = \begin{pmatrix} J_{1,1} & J_{1,2} & 0 \\ J_{1,2}' & J_{2,2} & J_{2,3} \\ 0 & J_{2,3}' & J_{3,3} \end{pmatrix}$$

Information Geometry*

Based upon the *Kullback-Leibler divergence*†, a measure of contrast between probability distributions.

$$\boldsymbol{D}(\boldsymbol{p}\|\boldsymbol{q}) = \boldsymbol{E}_{\boldsymbol{p}}\left\{\log\frac{\boldsymbol{p}(\mathbf{x})}{\boldsymbol{q}(\mathbf{x})}\right\}$$

Bregman distance in $\boldsymbol{\theta}$ based upon $\boldsymbol{\varphi}(\boldsymbol{\theta})$,

$$\boldsymbol{D}(\boldsymbol{\theta}^*\|\boldsymbol{\theta}) = \boldsymbol{\varphi}(\boldsymbol{\theta}) - \boldsymbol{\nabla}\boldsymbol{\varphi}(\boldsymbol{\theta}^*)\cdot(\boldsymbol{\theta}-\boldsymbol{\theta}^*)$$

Legendre transform $\boldsymbol{\varphi}^*(\boldsymbol{\eta})$ of $\boldsymbol{\varphi}(\boldsymbol{\theta})$:

$$\boldsymbol{\varphi}^*(\boldsymbol{\eta}) = \boldsymbol{\theta}(\boldsymbol{\eta})\cdot\boldsymbol{\eta} - \boldsymbol{\varphi}(\boldsymbol{\theta})$$

"Slope transform"

$$\begin{aligned}\boldsymbol{\eta}(\boldsymbol{\theta}) &= \frac{\boldsymbol{\partial}\boldsymbol{\varphi}(\boldsymbol{\theta})}{\boldsymbol{\partial}\boldsymbol{\theta}}\\ \boldsymbol{\theta}(\boldsymbol{\eta}) &= \frac{\boldsymbol{\partial}\boldsymbol{\varphi}^*(\boldsymbol{\eta})}{\boldsymbol{\partial}\boldsymbol{\eta}}\end{aligned}$$

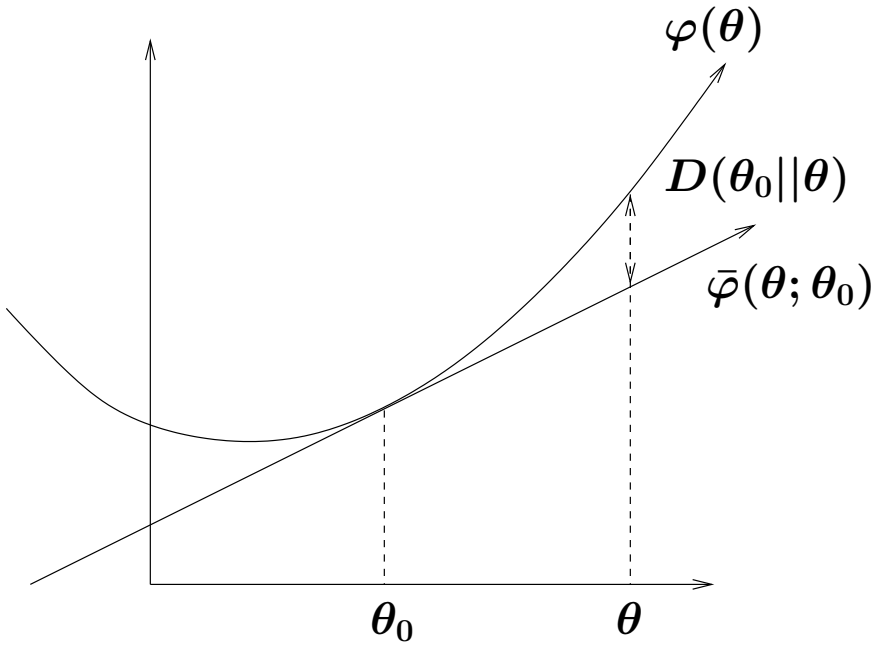
Convex bifunction in $(\boldsymbol{\eta}(\boldsymbol{p}),\boldsymbol{\theta}(\boldsymbol{q}))$,

$$\boldsymbol{D}(\boldsymbol{\eta}\|\boldsymbol{\theta}) = \boldsymbol{\varphi}^*(\boldsymbol{\eta}) + \boldsymbol{\varphi}(\boldsymbol{\theta}) - \boldsymbol{\eta}\cdot\boldsymbol{\theta}$$

*Chentsov, 72; Csiszár, 75; Efron, 78; Amari, 01.

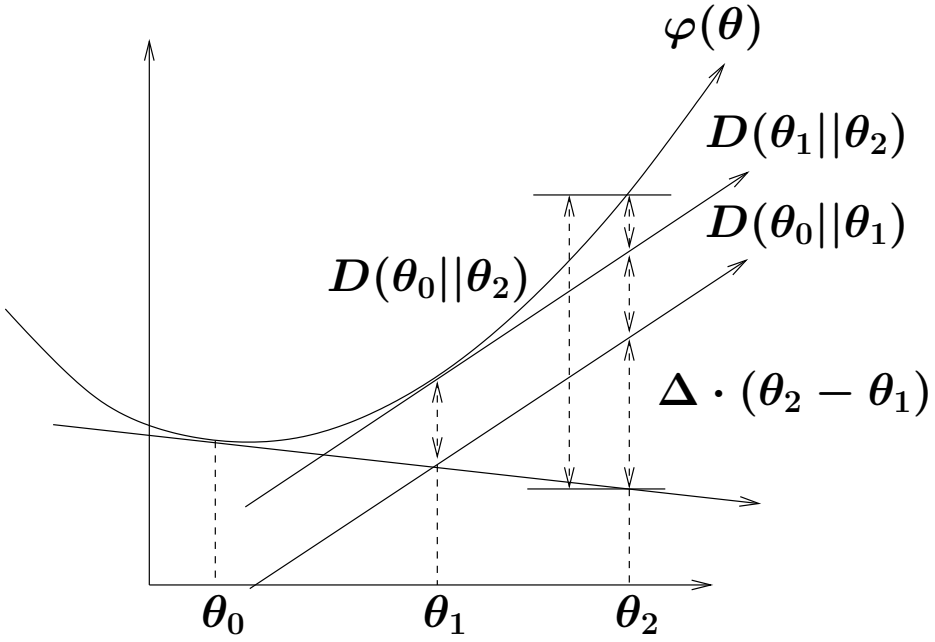
†Kullback and Leibler, 51.

Bregman distance*



*Bregman, 67.

Triangle Relation



$$D(\theta_0\|\theta_2) = D(\theta_0\|\theta_1)+D(\theta_1\|\theta_2)+(\eta_1-\eta_0)\cdot(\theta_2-\theta_1)$$

Information Projections

Let $\mathcal{F}$ be a regular exponential family with minimal statistics $t(x)$, exponential coordinates Θ, and moment coordinates $\eta(\Theta)$.

M-projection. Let $p \in \mathcal{F}$, $\mathcal{H} \subset \mathcal{F}$ e-flat submanifold. Exists unique $q^* \in \mathcal{H}$ satisfying the following equivalent conditions:

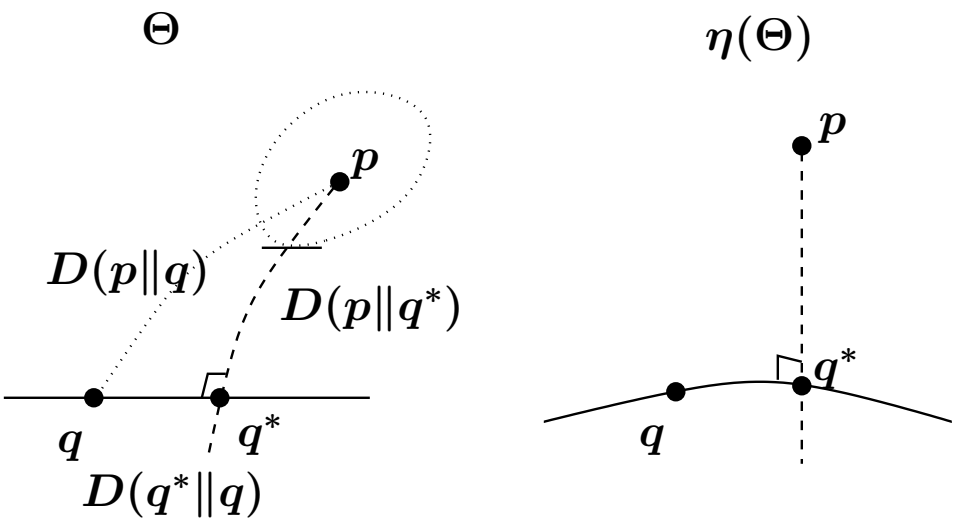
(i) $D(p\|q^*) = \inf_{q\in\mathcal{H}} D(p\|q)$

(ii) $\forall q \in \mathcal{H}: \ (\eta(p)-\eta(q^*))\cdot(\theta(q)-\theta(q^*)) = 0$

(iii) $\forall q \in \mathcal{H}: \ D(p\|q) = D(p\|q^*) + D(q^*\|q)$

We call $q^* = \arg\min_{q\in\mathcal{H}} D(p\|q)$ the *m-projection* of p to $\mathcal{H}$.

M-projection



$$\frac{\partial}{\partial \theta(q)} D(p\|q) = \eta(q) - \eta(p)$$

Dual E-projection

E-projection. Let $\boldsymbol{q} \in \boldsymbol{\mathcal{F}}$, $\boldsymbol{\mathcal{H}'} \subset \boldsymbol{\mathcal{F}}$ m-flat submanifold. Exists unique $\boldsymbol{p^*} \in \boldsymbol{\mathcal{H}'}$ satisfying the following equivalent conditions:

(i) $\boldsymbol{D(p^*\|q) = \inf_{p\in\mathcal{H}'} D(p\|q)}$

(ii) $\boldsymbol{\forall p \in \mathcal{H}' :\ (\eta(p)-\eta(p^*))\cdot(\theta(q)-\theta(p^*)) = 0}$

(iii) $\boldsymbol{\forall p \in \mathcal{H}' :\ D(p\|q) = D(p\|p^*) + D(p^*\|q)}$

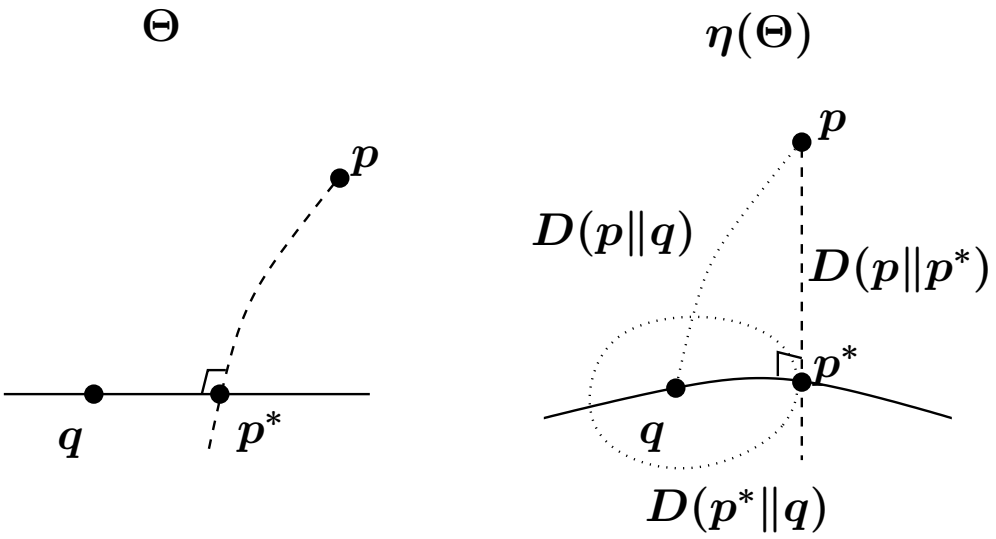
We call $\boldsymbol{p^* = \arg\min_{p\in\mathcal{H}'} D(p||q)}$ the *e-projection* of $\boldsymbol{q}$ to $\boldsymbol{\mathcal{H}'}$.

Duality. Let $\boldsymbol{\mathcal{H}}$ and $\boldsymbol{\mathcal{H}'}$ be *I-orthogonal* submanifolds such that exists $\boldsymbol{r}$ in intersection and

$\boldsymbol{\forall p \in \mathcal{H}', q \in \mathcal{H} : (\eta(p)-\eta(r))\cdot(\theta(q)-\theta(r)) = 0}$

Then, $\boldsymbol{r}$ is both the m-projection of $\boldsymbol{p} \in \boldsymbol{\mathcal{H}'}$ to $\boldsymbol{\mathcal{H}}$ and the e-projection of $\boldsymbol{q} \in \boldsymbol{\mathcal{H}}$ to $\boldsymbol{\mathcal{H}'}$.

E-projection



$$\frac{\partial}{\partial \eta(p)} D(p\|q) = \theta(p) - \theta(q)$$

Model Thinning

Let $t(x) = (t_{\mathcal{H}}(x), t'_{\mathcal{H}}(x))$, $\theta = (\theta_{\mathcal{H}}, \theta'_{\mathcal{H}})$ and $\eta = (\eta_{\mathcal{H}}, \eta'_{\mathcal{H}})$.

Objective. M-project $p \in \mathcal{F}$ to lower-order exponential family,

$$\mathcal{H} = \{q \in \mathcal{F} \mid \theta'_{\mathcal{H}}(q) = 0\}$$

Dual Problem. E-projection $q \in \mathcal{H}$ to the m-flat submanifold:

$$\mathcal{H}'(p) = \{r \in \mathcal{F} \mid \eta_{\mathcal{H}}(r) = \eta_{\mathcal{H}}(p)\} \qquad (1)$$

The latter e-projection problem may be solved by iterative scaling techniques which adjust parameters $\theta_{\mathcal{H}}(q)$ until $\eta_{\mathcal{H}}(q) = \eta_{\mathcal{H}}(p)$ (moment matching).

For GMRF $\mathbf{x} \sim \mathcal{N}^{-1}(h, J)$, impose sparsity on J. Moment-matching gives classical *covariance selection problem* (Dempster, 72).

Iterative Scaling

Alternating e-projections to set of m-flat submanifolds converges to e-projection to intersection (Csiszár, 75). Special case of method of alternating Bregman projections (Bregman, 67).

*Iterative Proportional Fitting.** m-flat submanifolds impose marginal moment constraints specifing marginal distribution $p^*(x_C)$.

$$\psi(x_C) \leftarrow \psi(x_C) \times \frac{p^*(x_C)}{p(x_C)}$$

Covariance Selection.† Updates exponential parameters (h_C, J_C) to impose moment constraints (μ^*_C, Σ^*_C).

$$\begin{aligned} J_C &\leftarrow J_C + (J^*_C - \hat{J}_C) \\ h_C &\leftarrow h_C + (h^*_C - \hat{h}_C) \end{aligned}$$

where $(h^*_C, J^*_C) = ((\Sigma^*_C)^{-1}\mu^*_C, (\Sigma^*_C)^{-1})$ and $(\hat{h}^*_C, \hat{J}^*_C) = (\Sigma_C^{-1}\mu_C, \Sigma_C^{-1})$ (marginal information models).

*Ireland and Kullback, 68.

†Speed and Kivveri, 86.

Greedy Edge-Removal

Prunes edges from graphical model by forcing selected off-diagonal entries of $\boldsymbol{J}$ to zero (m-projections implemented by iterative scaling techniques).

Selects weak interactions to prune according to conditional mutual information

$$\boldsymbol{I}(\mathbf{x}_i;\mathbf{x}_j|\mathbf{x}_{\setminus ij}) = -\frac{\mathbf{1}}{\mathbf{2}}\log\left(1-\frac{\det \boldsymbol{J}_{i,j}}{\sqrt{\det \boldsymbol{J}_{i,i}\det \boldsymbol{J}_{j,j}}}\right)$$

which gives tractable lower-bound estimate of KL under m-projection.

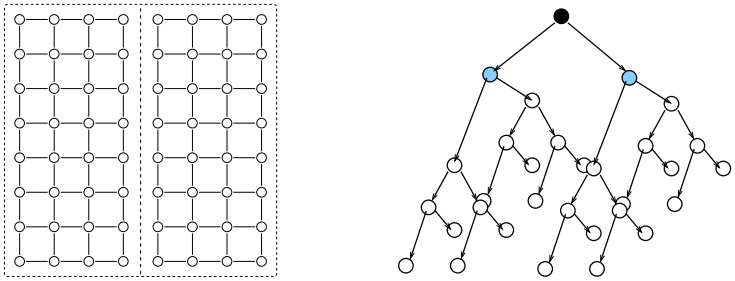
Selects batch $\boldsymbol{\mathcal{K}} \subset \boldsymbol{\mathcal{V}}$ of weakest edges to prune satisfying

$$\sum_{\boldsymbol{\mathcal{K}}} \boldsymbol{I}_{i;j} < \frac{\boldsymbol{\delta}}{|\boldsymbol{\mathcal{K}}|}$$

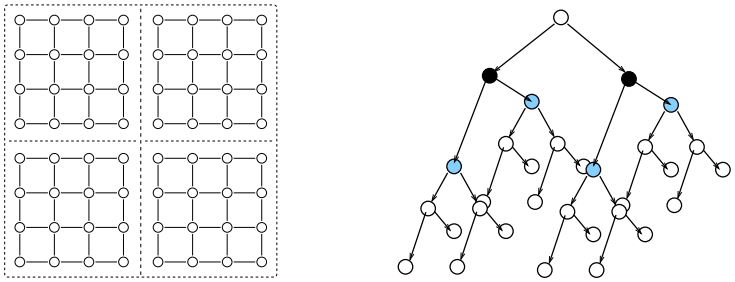
Continues thinning until no more weak interactions relative to $\boldsymbol{\delta}$. Related to Akaike information criterion (Akaike, 74).

Nested Dissection

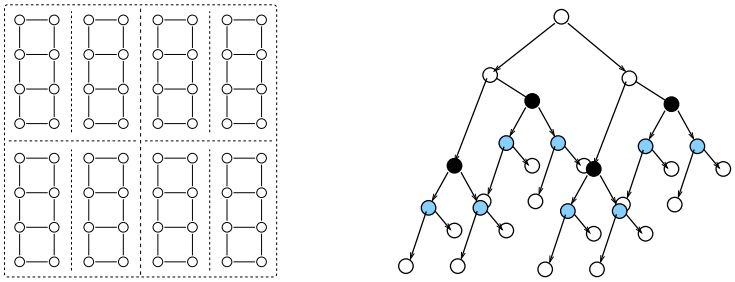
(1) vertical cut.



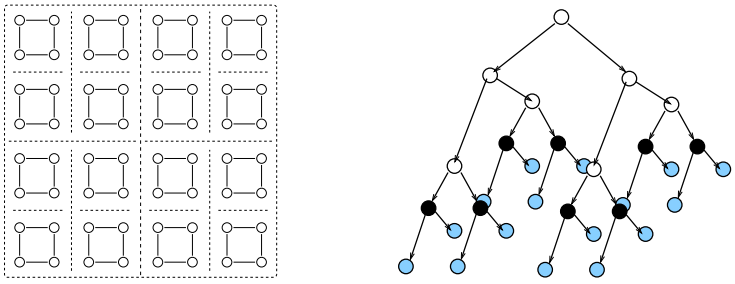
(2) horizontal cut.



(3) vertical cut



(4) horizontal cut.



Variable Elimination

Integrate over subset $\boldsymbol{\Lambda} \subset \boldsymbol{\mathcal{V}}$ of random variables:

$$\boldsymbol{p}(\boldsymbol{x}_{\setminus \boldsymbol{\Lambda}}) = \int \boldsymbol{p}(\boldsymbol{x}) \boldsymbol{dx}_{\boldsymbol{\Lambda}}$$

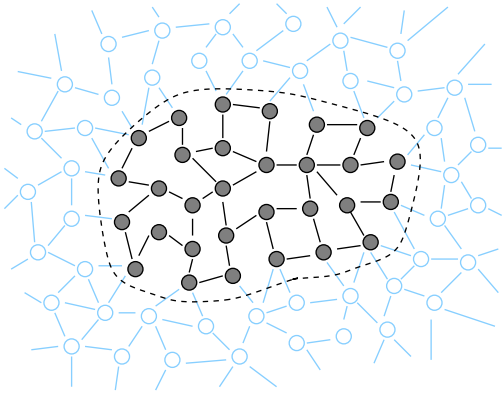
Local parameter update in $(\boldsymbol{h}, \boldsymbol{J})$ representation:

$$\begin{aligned} \boldsymbol{h}_{\boldsymbol{\partial\Lambda}} &\leftarrow \boldsymbol{h}_{\boldsymbol{\partial\Lambda}} - \boldsymbol{J}_{\boldsymbol{\partial\Lambda},\boldsymbol{\Lambda}} \boldsymbol{J}_{\boldsymbol{\Lambda},\boldsymbol{\Lambda}}^{-1} \boldsymbol{h}_{\boldsymbol{\Lambda}} \\ \boldsymbol{J}_{\boldsymbol{\partial\Lambda}} &\leftarrow \boldsymbol{J}_{\boldsymbol{\partial\Lambda}} - \boldsymbol{J}_{\boldsymbol{\partial\Lambda},\boldsymbol{\Lambda}} \boldsymbol{J}_{\boldsymbol{\Lambda},\boldsymbol{\Lambda}}^{-1} \boldsymbol{J}_{\boldsymbol{\Lambda},\boldsymbol{\partial\Lambda}} \end{aligned}$$

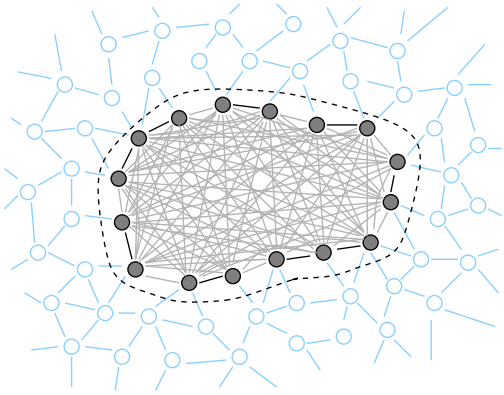
Eliminates vertices in graphical model but adds "fill" edges between neighbors. Only updates local parameters and structure of "boundary" $\boldsymbol{\partial\Lambda}$ of subfield.

Cavity Models (Initialization)

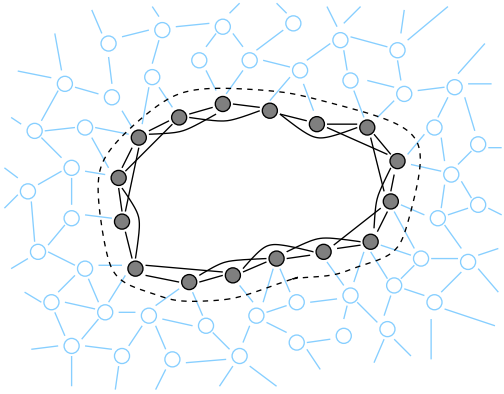
(1) Partial model of subfield (zero boundary).



(2) Elimination gives model of surface.

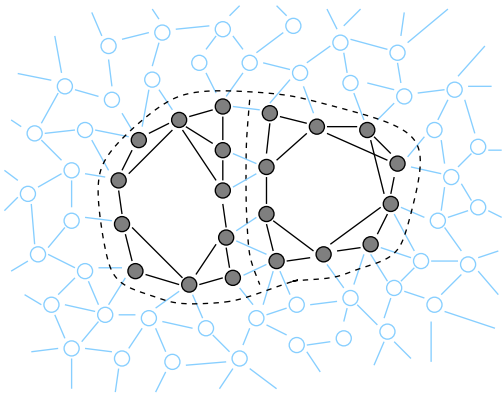


(3) Model thinning gives "cavity model".

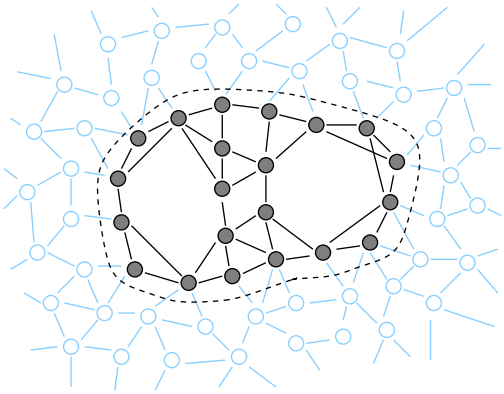


"Upwards" Cavity Modeling

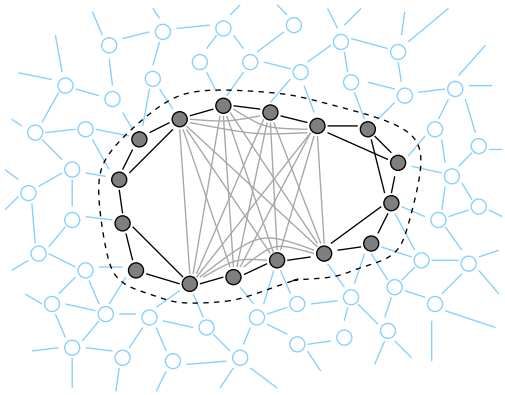
(1) Initialization.



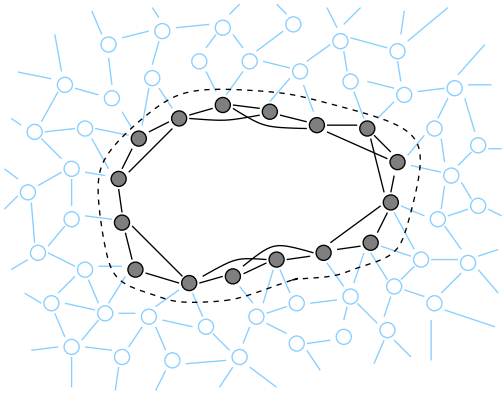
(2) Merge.



(3) Eliminate.

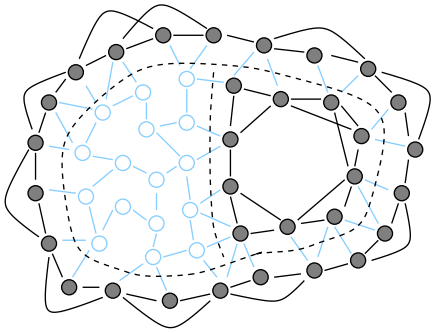


(4) Thin.

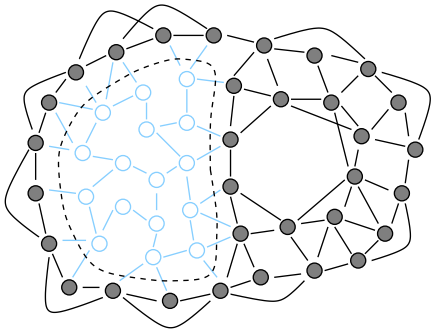


"Downwards" Blanket Modeling

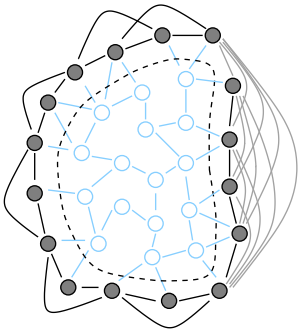
(1) Initialization.



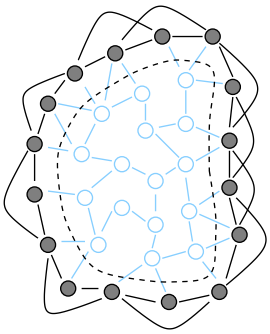
(2) Merge.



(3) Eliminate.



(4) Thin.



Conclusion

RCM appears to provide a powerful and flexible framework for tractable yet near-optimal computation in MRFs.

Much work remains to better characterize performance and explore promising extensions:

- Develop information geometry of RCM.
- Consider more general families of graphical models.
- Employ alternative modeling techniques.
- Applications
  - Model Identification
  - Image Processing
  - Data Compression and Coding
  - Monte-Carlo Simulation

References

Akaike, 74. A new look at the statistical model identification. *IEEE Trans. Auto. Control*, AC-19:716:723.

Amari, 01. Information geometry of hierarchy of probability distributions. *IEEE Trans. Inf. Theory*, 47(5):1701-1711.

Chentsov, 66. A systematic theory of exponential families. *Theory of Prob. and Appl.*, 11.

Chentsov, 72. Statistical decision rules and optimal inference. *AMS Trans. Math. Mono.*, v.53 (reprint 82).

Barndorff-Nielsen, 78. *Information and Exponential Families*. John Wiley.

Bregman, 67. The relaxation method of finding the common point of convex sets. *USSR Comp. Math. and Physics*, 7:200-217.

Csiszár, 75. I-divergence geometry of probability distributions and minimization problems. *Annals of Prob.*, 3(1):146-158.

Dempster, 72. Covariance Selection. *Biometrics,* 28(1):157-175.

Efron, 78. The geometry of exponential families. *Annals of Stat.*, 6(2):362-376.

Grimmett, 73. A thoerem about random fields. *Bull. of London Math. Soc.*, 5:81-84.

Ireland and Kullback, 68. Contingency tables with given marginals. *Biometrika*, 55:179-188.

Jordan (editor), 99. *Learning in Graphical Models.* MIT Press.

Kullback and Leibler, 51. On information and sufficiency. *Annals of Math. Stat.*, 22(1):79-86.

Lauritzen, 96. *Graphical Models*. Oxford University Press.

Speed and Kiiveri, 86. Gaussian Markov distributions over finite graphs. *Annals of Stat.*, 14(1):138-150.