# Nonlinear Optimization of Exponential Family Graphical Models 6.252 Term Project – Spring 2002

Ayres Fan, Jason K. Johnson, Dmitry M. Malioutov

May 17, 2002

# Overview

- Problem Statement
- Background Theory
  - Exponential Family and Information Geometry
  - Graphical Models
  - Gauss-Markov Processes
- Iterative M-Projection Algorithms
  - Gradient and Hessian Calculations
  - Nonlinear Optimization Techniques
- Structure Estimation via M-Projections & AIC/BIC

## Problem Statement

*Moment Matching.* Concerns a regular exponential family of models for a random variable $\mathrm{x}$ having exponential parameters $\boldsymbol{\theta}$, minimal set of sufficient statistics $\boldsymbol{t}(\boldsymbol{x})$ and base measure $\boldsymbol{q}(\boldsymbol{x}) > \mathbf{0}$ with pdf

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

A dual parameterization of this family is given by the moment coordinates

$$\boldsymbol{\eta} = \boldsymbol{E}_{\boldsymbol{\theta}}\{\boldsymbol{t}(\mathrm{x})\}$$

which are in one-to-one correspondence with exponential coordinates.

The moment-matching problem is to to recover $\boldsymbol{\theta}$ given $\boldsymbol{\eta}$.

Solve $\boldsymbol{\eta}(\boldsymbol{\theta}) = \boldsymbol{\eta}^*$.

## The Exponential Family*

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 $\Theta$ s.t. pdf is normalizable $\varphi(\theta) < \infty$. The family is *regular* if $\Theta$ 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 $\eta(\Theta)$.

*Efron, 78; Barndorff-Nielsen, 1978.

• *Maximum Entropy Principle.*[*] The probability density function $p(x)$ which has maximum entropy

$$h[p] = -\int p(x) \log p(x) dx$$

subject to moment constraints

$$\int p(x) t(x) dx = \eta^*$$

is an exponential family family with $q(x) = 1$ and statistics $t(x)$ where $\theta^*$ is determined by condition $\eta(\theta^*) = \eta^*$.

• *Minimum Relative-Entropy Principle.*[†] Given a reference model $q(x)$, the density $p(x)$ so as to minimize the Kullback-Leibler divergence[‡]

$$D(p||q) = \int p(x) \log \frac{p(x)}{q(x)} dx$$

again subject to moment constraints $E_p\{t(x)\} = \eta^*$ is an exponential family model with base measure $q(x)$ and statistics $t(x)$ where $\theta^*$ is determined by $\eta(\theta^*) = \eta^*$.

[*] Jaynes, 58; and Good,63.

[†] Kullback and Leibler, 51.

[‡] aka relative or cross entropy as is invariant form of entropy.

## Relation to Maximum Likelihood (ML)

Latter "KL-projection"* arises in ML parameter estimation. Given $x^{(1)}, \ldots, x^{(N)} \sim p(\mathrm{x})$, the member of a given exponential family which maximizes the joint log-likelihood of the data

$$\hat{\theta}_{ML} = \arg\max_{\theta} \sum_{k=1}^{N} \log f(x^{(k)}; \theta)$$

is determined by KL-projection as it minimizes $D(\tilde{p} || f(\cdot; \theta))$ where $\tilde{p}(x)$ is the empirical distribution

$$\tilde{p}(x) = \sum_{k=1}^{N} \delta(x - x^{(k)})$$

having the same moments as the data

$$E_{\tilde{p}} t(\mathrm{x}) = \tilde{\eta} = \frac{1}{N} \sum_{k=1}^{N} t(x^{(k)})$$

The maximum-likelihood parameters may then be determined by moment-matching $\eta(\theta) = \tilde{\eta}$.

*Csiszár, 75; Amari, 01.

## Graphical Models*

Consider an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with $\mathcal{V}$ denoting the set of vertices of the graph and $\mathcal{E}$ denoted the set of edges. Let $\mathcal{V}$ index elements of $\mathrm{x}$. Then, $\mathrm{x}$ is said to be *Markov* w.r.t $\mathcal{G}$ if for each vertex $i$ the state $\mathrm{x}_i$ is conditionally independent of all non-neighbors given the state of just the neighbors $j \in \mathcal{V} : \langle ij \rangle \in \mathcal{E}$.

The *Hammersley-Clifford theorem* states that $\mathrm{x}$ is Markov w.r.t. $\mathcal{G}$ if and only if pdf factors according to $\mathcal{G}$ as

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

where *potentials* are positive compatibility functions and $Z(\psi)$ is just a normalization constant.

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

*Lauritzen, 96; Jordan, 99.

## Exponential Family Graphical Models

Restrict statistics $\boldsymbol{t}(\boldsymbol{x})$ to consist solely of "local" statistics on cliques of vertices $\boldsymbol{t_c}(\boldsymbol{x_c})$ then the exponential family pdf given earlier factors as above with potential functions

$$\boldsymbol{\psi_c}(\boldsymbol{x_c}) = \exp\{\boldsymbol{\theta_c} \cdot \boldsymbol{t_c}(\boldsymbol{x_c})\}$$

and normalization constant

$$\boldsymbol{Z}(\boldsymbol{\psi}) = \exp\{\boldsymbol{\varphi}(\boldsymbol{\theta})\}$$

and is thus Markov w.r.t. $\boldsymbol{\mathcal{G}}$.

Includes all $\boldsymbol{\mathcal{G}}$-Markov processes which may be parameterized s.t. log-potentials vary linearly in the parameters.

# Gaussian Processes

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

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

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

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

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

## Gaussian Hammersley-Clifford

Suppose $\mathbf{x} \sim \mathcal{N}^{-1}(h, J)$ is $\mathcal{G}$-Markov.

The *partial correlation coefficients**

$$\rho(\mathbf{x}_i, \mathbf{x}_j | \mathbf{x}_{ij}^c) = \frac{\mathrm{cov}(\mathbf{x}_i, \mathbf{x}_j | \mathbf{x}_{ij}^c)}{\sqrt{\mathrm{cov}(\mathbf{x}_i | \mathbf{x}_{ij}^c)\mathrm{cov}(\mathbf{x}_j | \mathbf{x}_{ij}^c)}}$$

related to *conditional mutual information*

$$I(\mathbf{x}_i; \mathbf{x}_j | \mathbf{x}_{ij}^c) = -\frac{1}{2}\log(1 - \rho^2(\mathbf{x}_i, \mathbf{x}_j | \mathbf{x}_{ij}^c))$$

readily evaluated as

$$\rho(\mathbf{x}_i, \mathbf{x}_j | \mathbf{x}_{ij}^c) = -\frac{J_{ij}}{\sqrt{J_{ii}J_{jj}}}$$

$\mathcal{G}$-Markov property satisfied if and only if PCC and MI are zero for non-edges s.t. $J$ has same sparsity structure as $\mathcal{G} = (\mathcal{V}, \mathcal{E})$.

$$J_{ij} \neq 0 \Leftrightarrow \langle ij \rangle \in \mathcal{E}$$

Information filter form $(h, J)$ provides compact graphical model with $J$ sparse.

*Lauritzen, 96.

# Gauss-Markov Process Exponential Description

- Statistics of $\mathrm{x}$ on $\mathcal{G} = (\mathcal{V}, \mathcal{E})$

$$t_\gamma(x) = \begin{cases} (x_i, x_i^2), & \gamma = i \in \mathcal{V} \\ x_i x_j, & \gamma = \langle ij \rangle \in \mathcal{E} \end{cases}$$

- Parameters $\theta \Leftrightarrow (h, J)$

$$\theta_\gamma = \begin{cases} (h_i, -J_{ii}/2), & \gamma = i \in \mathcal{V} \\ -J_{ij}, & \gamma = \langle ij \rangle \in \mathcal{E} \end{cases}$$

- Moments $(\mu, \Sigma) \Rightarrow \eta$

$$\eta_\gamma = \begin{cases} (\mu_i, \Sigma_{ii} + \mu_i^2), & \gamma = i \in \mathcal{V} \\ \Sigma_{ij} + \mu_i \mu_j, & \gamma = \langle ij \rangle \in \mathcal{E} \end{cases}$$

- "Brute force" inference of $\eta(\theta)$ performs $(h, J) \Rightarrow (\mu, \Sigma)$ by

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

so that $\theta \Rightarrow (h, J) \Rightarrow (\mu, \Sigma) \Rightarrow \eta$. Note that $(\mu, \Sigma)$ not fully specified by $\eta$ such that moment-matching is nontrivial.

## Iterative Methods for M-Projection

Pose moment-matching as "m-projection" KL-projection to exponential family $\mathcal{F}$.

$$\begin{array}{lll} \textbf{(P)} & \textbf{minimize} & D(\eta^*||\theta) \\ & \textbf{s.t.} & \theta \in \Theta \end{array}$$

where $\eta^* \in \eta(\Theta)$, $D(\eta^*||\theta)$ is KL-divergence between (unknown) density $f^* \in \mathcal{F}$ with moments $\eta^*$ and $f(\cdot;\theta)$. KL-divergence may be expressed as

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

where the cumulant function $\varphi(\theta)$ and its convex conjugate $\varphi^*(\eta) = \sup_\theta\{\varphi(\theta) - \theta \cdot \eta\}$ (negative entropy) are strictly* convex functions such that $D(\eta^*||\theta)$ is convex in either argument. Admissable parameter set $\Theta$ is convex. Convex programming problem, equivalent to minimizing $g(\theta) = \varphi(\theta) - \eta^* \cdot \theta$. Related to "barrier" method of semi-definite programming problem for Gaussian family.

*under regularity and minimality assumptions

## Gradient and Hessian of KL-divergence.

Gradient of the cumulant function generates the moments.

$$\nabla_{\theta}\varphi(\theta) = \eta(\theta)$$

Hessian of the cumulant function generates the *Fisher information matrix* defined as the covariance of the sufficient statistics.

$$\begin{aligned}\nabla^2_{\theta}\varphi(\theta) &= G(\theta) \\ &= \mathrm{cov}_{\theta}(t(\mathrm{x})) \\ &= E_{\theta}\{t(\mathrm{x})t(\mathrm{x})'\} - \eta\eta'\end{aligned}$$

Consequently, the gradient of KL is just difference in moments

$$\nabla_{\theta}D(\eta^*||\theta) = \eta(\theta) - \eta^*$$

while the Hessian is the Fisher information

$$\nabla^2_{\theta}D(\eta^*||\theta) = G(\theta)$$

Evaluate the gradient using "brute force" inference described earlier.

# Evaluation of Fisher Information

For zero-mean Gaussian $\tilde{\mathrm{x}} \equiv \mathrm{x} - \boldsymbol{\mu}$ 3rd order moments are zero

$$E\{\tilde{\mathrm{x}}_i \tilde{\mathrm{x}}_j \tilde{\mathrm{x}}_k\} = 0$$

while the 4th order moments are given by 2nd order moments

$$E\{\tilde{\mathrm{x}}_i \tilde{\mathrm{x}}_j \tilde{\mathrm{x}}_k \tilde{\mathrm{x}}_l\} = \Sigma_{ij}\Sigma_{kl} + \Sigma_{ik}\Sigma_{jl} + \Sigma_{il}\Sigma_{jk}$$

Consequently, we arrive at the following formulas for the elements of $G(\boldsymbol{\theta})$.

$$\begin{aligned}
G_{i;j} &\equiv \mathrm{cov}(\mathrm{x}_i; \mathrm{x}_j) \\
&= \Sigma_{ij} \\
G_{ij;k} &\equiv \mathrm{cov}(\mathrm{x}_i \mathrm{x}_j; \mathrm{x}_k) \\
&= \Sigma_{ik}\mu_j + \Sigma_{jk}\mu_i \\
G_{ij;kl} &\equiv \mathrm{cov}(\mathrm{x}_i \mathrm{x}_j; \mathrm{x}_k \mathrm{x}_l) \\
&= \Sigma_{ik}\Sigma_{jl} + \Sigma_{il}\Sigma_{jk} + \Sigma_{ik}\mu_j\mu_l \\
&\quad + \Sigma_{il}\mu_j\mu_k + \Sigma_{jk}\mu_i\mu_l + \Sigma_{jl}\mu_i\mu_k
\end{aligned}$$

Evaluate sparse subset, e.g. $G_{\langle ij \rangle;k}, G_{ii;k}$.

Also requires "inference" computation

$$\boldsymbol{\theta} \Rightarrow (h, J) \Rightarrow (\boldsymbol{\mu}, \Sigma)$$

# Optimization Techniques

We perform minimization of $\varphi(\boldsymbol{\theta}) - \boldsymbol{\eta}^* \cdot \boldsymbol{\theta}$ employing earlier gradient $\boldsymbol{g}(\boldsymbol{\theta})$ and Hessian $\boldsymbol{G}(\boldsymbol{\theta})$ evaluators and the following standard* methods. All methods are initialized by m-projection of $\boldsymbol{\eta}^*$ to "fully factorized" (disconnected) family.

$$\boldsymbol{\theta}_\gamma^{(0)} = \begin{cases} (\boldsymbol{\mu_i}/\boldsymbol{\Sigma_{ii}}, \mathbf{1}/\boldsymbol{\Sigma_{ii}}), & \gamma = \boldsymbol{i} \in \boldsymbol{\mathcal{V}} \\ \mathbf{0}, & \gamma = \langle \boldsymbol{ij} \rangle \in \boldsymbol{\mathcal{E}} \end{cases}$$

*Gradient Descent.* line-minimization implemented by seeking zero of gradient along search direction (exploiting strict convexity). This is m-projection to e-geodesic.

*Conjugate Gradients.* uses "non-jamming" direction update and performs conjugacy test for early "restarts" with threshold $\mathbf{0.05}$.

*Preconditioned Conjugate Gradients.* as above with preconditioning matrix $\boldsymbol{M}$ chosen as either the inverse diagonal $\boldsymbol{M} = \boldsymbol{Diag}(\boldsymbol{G}(\boldsymbol{\theta}))^{-1}$ or as "full" inverse $\boldsymbol{M} = \boldsymbol{G}(\boldsymbol{\theta})^{-1}$.

*Newton's Method.* without line-minimization.

*Bertsekas, 95.

## Application to ML Estimation

Experiments examine performance of these methods for ML estimation of parameters of Gauss-Markov process from observed sample paths.

1. Construct "truth" model $(\mathcal{G}, \theta_{\mathbf{true}})$.

2. Generate sample-paths $x^{(1)}, \ldots, x^{(N)} \sim p(x)$ by Monte-Carlo simulation.

3. Sample-average statistics $\tilde{\eta} = \frac{1}{N} \sum_k t(x^{(k)})$.

4. Given $(\mathcal{G}, \tilde{\eta})$, iteratively solve $\eta(\theta) = \tilde{\eta}$.

Then, solution $\theta^*$ is ML-estimate of $\theta_{\mathbf{true}}$.

We generate truth models for testing with a variety of graphical structures (k-th order chains and loops, 2d nearest-neighbor grids, and random graphs) and generate random model $(h, J)$.

## M-Projections for Structure Estimation

Here we consider the case where the Markov structure $\mathcal{G}$ is unknown and we wish to provide a compact yet faithful model for the data by also estimating $\mathcal{G}$.

Employ either AIC* or BIC† to resolve trade-off between fitting the data and minimizing the complexity of the model $K_{\mathcal{G}} = 2 * |\mathcal{V}_{\mathcal{G}}| + |\mathcal{E}_{\mathcal{G}}|$.

$$\begin{array}{rc} \textbf{minimize} & D(\tilde{\eta}^*_{\mathcal{F}} || \theta_{\mathcal{G}}) + \delta K_{\mathcal{G}} \\ \textbf{w.r.t} & (\mathcal{G}, \theta_{\mathcal{G}}) \end{array}$$

where $\delta$ is specified threshold and $\mathcal{F}$ denotes the "full" graph so that $\tilde{\eta}^*_{\mathcal{F}} = (\tilde{\mu}, \tilde{\Sigma})$. For a trial $\mathcal{G}$ (having edges removed) the best $\theta_{\mathcal{G}}$ is given by m-projection $\theta^*_{\mathcal{G}}$ solving $\eta_{\mathcal{G}}(\theta_{\mathcal{G}}) = \tilde{\eta}^*_{\mathcal{G}}$ maintaining a subset of moment constraints.

*Pythagorean theorem*‡ If $\mathcal{G}1 \subset \mathcal{G}2 \subset \mathcal{F}$ then KL-divergence decomposes as

$$D(\theta_{\mathcal{F}} || \theta_{\mathcal{G}1}) = D(\theta_{\mathcal{F}} || \theta^*_{\mathcal{G}2}) + D(\theta^*_{\mathcal{G}2} || \theta_{\mathcal{G}1})$$

*Akaike, 74.

†Schwarz, 78.

‡Amari, 01.

## Greedy Algorithm

Pythagorean theorem suggests successive projections to embedded graphs having 1 less edge until KL-divergence exceeds $\delta$. Avoids combinatorial search over $\mathcal{G}$ but not necessarily optimal.

May greatly reduce search over embedded graphs by employing lower-bound

$$I_{\langle ij \rangle} \equiv I(x_i; x_j | x_{ij}^c) \leq D(\theta_{\mathcal{G}} || \theta^*_{\mathcal{G} \setminus \langle ij \rangle})$$

to eliminate strong interactions from consideration for projection. Lower-bound is easily calculated as,

$$I_{\langle ij \rangle} = -\frac{1}{2} \log \left( 1 - \frac{J_{ij}^2}{J_{ii} J_{jj}} \right)$$

*Outline.* Starting with $(\mathcal{F}, \tilde{\eta})$, performs successive m-projections to lower-order embedded graph having one less edge. Select edge to prune as follows:

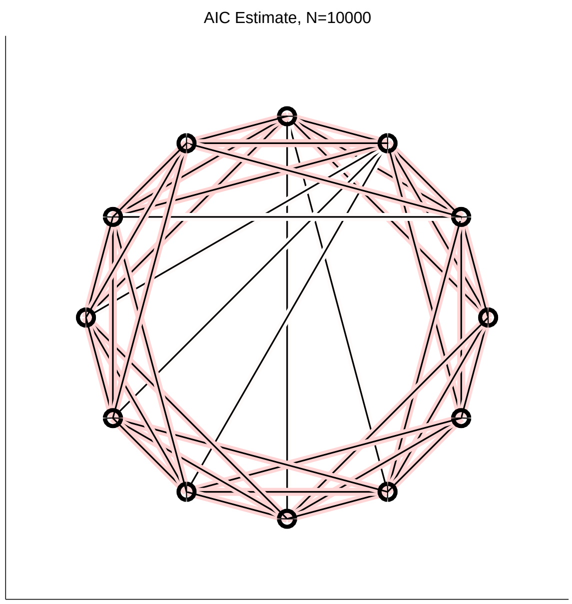
1. Identify candidate edges with $I_{\langle ij \rangle} < \delta$.

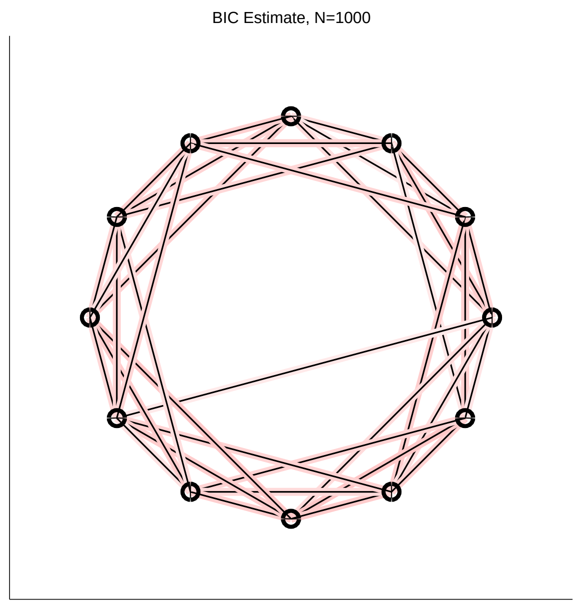
2. For lowest $I_{\langle ij \rangle}$ candidate, project to $\mathcal{G} \setminus \langle ij \rangle$.

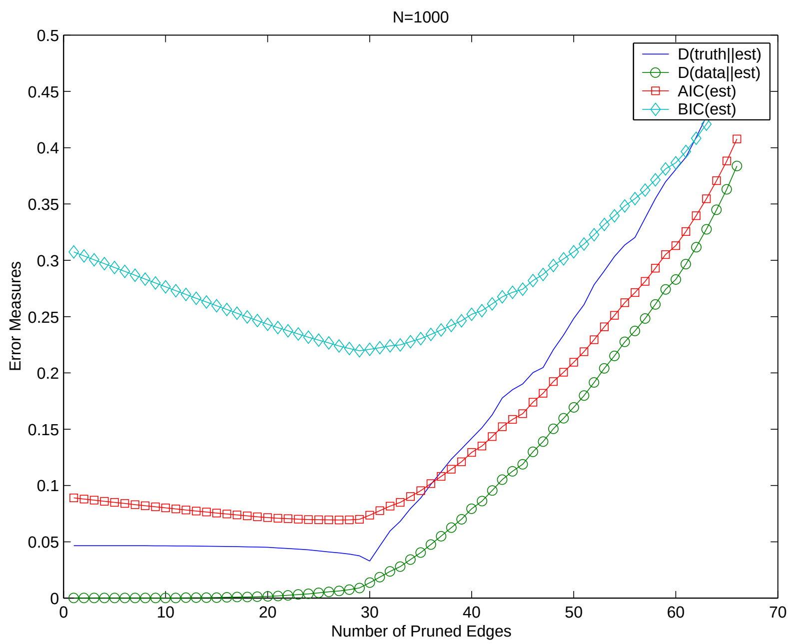
3. If improvement, update $(\mathcal{G}, \theta^*_{\mathcal{G}})$ and eliminate any candidates worst then observed KL-divergence.

4. While untried candidates remain, goto (1).

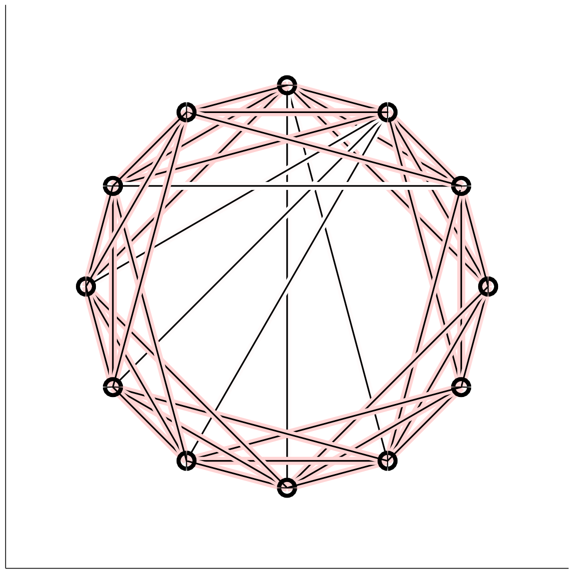
When all candidates have been checked or eliminated, the best candidate is accepted if KL-divergence less than $\delta$. Otherwise, terminates with current estimate.


AIC Estimate, N=10000

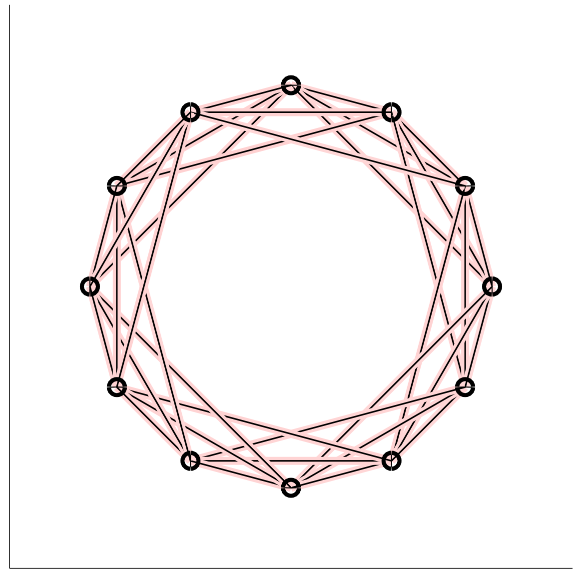

BIC Estimate, N=1000

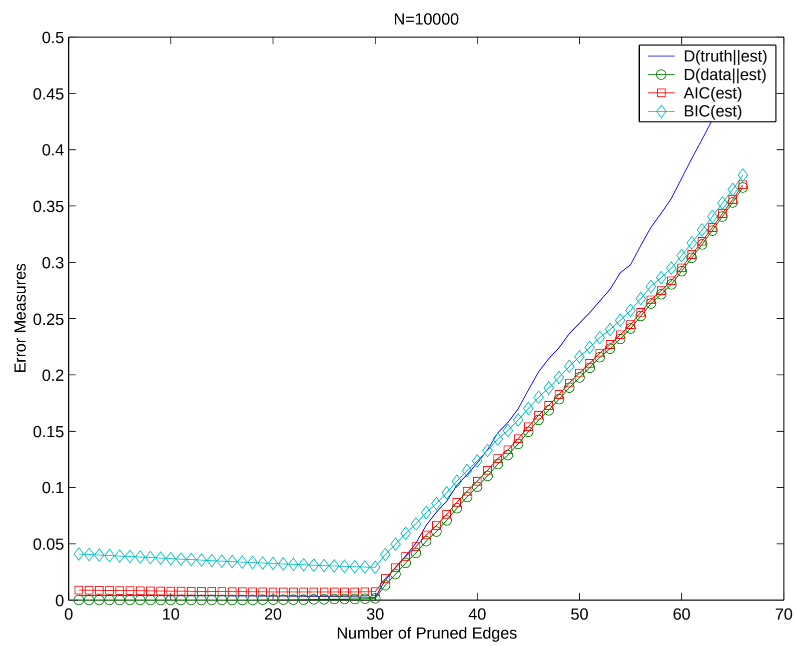

N=1000
D(truth||est)
D(data||est)
AIC(est)
BIC(est)
Error Measures
Number of Pruned Edges
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.4
0.45
0.5
0
10
20
30
40
50
60
70

AIC Estimate, N=10000



BIC Estimate, N=10000





# Conclusions

- Moment-matching/M-Projection is well-posed convex problem.

- Standard optimization techniques work quite well and are robust.

- Outperforms standard Iterative Proportional Fitting (coordinate descent) approach.

- Newton's method most efficient for small graphs.

- Conjugate Gradients and Diagonal PCG more appropriate for larger problems provided efficient inference is available.

- Enables structure estimation with AIC/BIC.

## 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.

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

Bertsekas, 95. *Nonlinear Programming*. Athena Scientific.

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

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

Good, 63. Maximum entropy for hypothesis formulation. *Annals of Math. Stat.*, 34(3):911-934.

Jaynes, 57. Information theory and statistical mechanics. *Phys. Review*,106:620-630.

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.

Schwarz, 78. Estimating the dimension of a model. *Annals of Stat.*, 6:461-464.