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

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

May 17, 2002

***Abstract.*** *This project explores methods for carrying out projections arising in the information geometry of the exponential family of probability models. Kullback-Leibler divergence serves as the distance measure between probability models in this context. The applications include maximum likelihood parameter estimation given sample paths of an unknown density as well as model reduction where one wishes to fit a lower-order exponential model to a given higher-order model. These are fundamental problems arising in the context of the graphical modeling literature. Here, one considers exponential family models defined on graphs such that the random process is constrained to be Markov with respect to some interaction graph. In this context, we will show that the fundamental problem of minimizing the Kullback-Leibler divergence may be reduced to a certain moment-matching problem which may be posed as a convex programming problem. This minimization problem is then well-suited to solution by a variety of generic nonlinear programming methods. We will explore these methods in the context of Gaussian processes which are Markov with respect to a given graph and employ both gradient and Hessian based techniques (gradient descent, conjugate gradients, preconditioned conjugate gradients and Newton's method). We compare the performance of these methods to the standard iterative proportional fitting method for solving this moment-matching problem. Also, we extend these methods to explore the problem of structure estimation employing either the Akaike or Bayesian Information Criterion to estimate the graphical structure of a Gaussian process from data without any prior knowledge of the Markov structure of the process.*

## 1 Introduction

In this paper we develop general nonlinear optimization techniques for exponential family graphical models and implement these techniques for Gaussian processes. In the general context of exponential family models we consider the following fundamental moment-matching problem. Consider an exponential family model [Efr78, BN78] for a continuous-valued random vector x specified by a parameterized probability density function $f(x;\theta) = \exp\{\theta \cdot t(x) - \varphi(\theta)\}$ where $t(x)$ is a specified vector of statistics of $x$ and $\varphi(\theta)$ is an appropriately chosen normalization constant for each setting of the model parameters $\theta$. The *moments* of this model are the expected values of the statistics $\eta \equiv E_\theta t(\mathrm{x})$ (here $E_\theta\{\cdot\}$ denotes expectation with respect to the density $f(\cdot;\theta)$). Under certain regularity and minimality assumptions with respect to the chosen exponential family, it may be shown that there exists a one-to-one correspondence between model parameters $\theta$ and moments $\eta$ (over the set of moment values achievable by that family). The moment-matching problem may then be posed as follows. Suppose that we specify a desired value for the moments $\eta^*$ (contained in this achievable set) and we wish to determine the value of $\theta$ which achieves those moments. That

is we wish to solve the nonlinear equation $\eta(\theta) = \eta^*$. This is the basic moment-matching problem addressed by this paper. We assume throughout the "inference" computation of evaluating the moments $\eta(\theta)$ given $\theta$ is tractable and available as a subroutine to the iterative methods to be developed.

This fundamental problem arises in many contexts such as maximum entropy, minimum relative-entropy, and maximum likelihood modeling. These types of model-selection problems are especially important and challenging in the context of graphical modeling where the random process is presumed to have some Markov structure. This structure is advantageous in that it simplifies the representation of the model by reducing the number of model parameters required to describe the process. Much of the following discussion applies to the general case of exponential family models having such Markov structure. But for the sake of concreteness, we will focus upon the family of Gauss-Markov random fields which is a special case of exponential family graphical models. Yet, this is probably the most important case in practice. Also, we emphasize the generality of the approach in that our methods could be applied to many other exponential family graphical models.

In the next section we present the necessary background theory to correctly pose the moment-matching problem and appreciate its fundamental importance as well as to provide the necessary machinery to support the optimazation methods to be developed. Section 3 then details those nonlinear programming techniques we have applied to the moment-mathcing problem. The performance of these methods is examined in the context of maximum-likelihood parameter estimation of a Gauss-Markov random field (having known Markov structure) from sample paths of the process. Finally, these methods are applied to the problem of also estimating the Markov structure of the field from data. In order to resolve the trade-off between model fidelity and model complexity we employ either the Akaike Information Critierion (AIC) or the Bayesian Information Criterion (BIC) to select the order of the estimated graphical model.

## 2 Background Theory

Here we give a brief overview of the pertinent theory of the information geometry of exponential family models and illustrate the connection to graphical modeling with emphasis on the Gaussian case.

### 2.1 The Exponential Family and Information Geometry

An exponential family of models is specified by a positive function $q(x)$ (often taken to be unity for all $x$) which we refer to as the *base measure* and a set of *statistics* $t(x)$ both defined over some specified state-space $X$. In the case that $X$ is taken to be $R^n$, then $x$ is a continuous-valued random vector having probability density function of the form

$$f(x;\theta) = q(x)\exp\{\theta \cdot t(x) - \varphi(\theta)\} \tag{1}$$

minimalwhere $\varphi(\theta)$ is the normalization constant below sometimes referred to as the *cumulant function.*

$$\varphi(\theta) = \log \int q(x)\exp\{\theta \cdot t(x)\}dx \tag{2}$$

The corresponding exponential family is then given by the set of such densities which are normalizable such that the above normalization constant is finite. The model parameters $\theta$ are allowed to vary over only those admissible settings such the density is normalizable and are then considered as the exponential coordinates of this family of densities. An exponential family is said to be

*regular* when the set of admissible $\theta$ has non-empty interior. A regular exponential family is then said to be *minimal* when the statistics are linearly independent[1]. We shall assume throughout the remainder of this paper that we are concerned with representations of regular exponential families. Such a family has the important property that the set of admissible exponential coordinates $\theta$ are in one-to-one correspondence with the so-called moment coordinates $\eta = E_\theta t(x)$. In principle, the density is equally well specified by its moments $\eta$ as by its parameters $\theta$. Yet, recovering $\theta$ for a given $\eta$ often proves to be a difficult problem which can only be solved by iterative methods. This is the fundamental moment-matching problem which concerns us here.

*Maximum Entropy.* [Jay57, Goo63] To appreciate the fundamental nature of the moment matching problem, consider the maximum-entropy modeling paradigm. Here, one seeks the probability density function $p(x)$ (not presumed to be an exponential family model) which has maximum entropy

$$h[p] = -\int p(x) \log p(x) dx \tag{3}$$

subject to a set of moment constraints of the form

$$E_p t(\mathrm{x}) \equiv \int p(x) t(x) dx = \eta^* \tag{4}$$

where $t(x)$ are some specified set of statistics. Employing the method of Lagrange multipliers, it may be shown that the optimal density then has the form of an exponential family model with statistics $t(x)$ and base measure $q(x) = 1$ so that $p^*(x) = f(x;\theta^*)$ where the model parameters $\theta$ arise as the Lagrange multipliers and are determined by the condition that $\theta^*$ solves $\eta(\theta) = \eta^*$. Thus, we have illustrated the maximum-entropy interpretation of the exponential family models and recovered the moment-matching problem in this context.

*Minimum Relative Entropy.* [KL51, Kul59] The minimum relative entropy modeling principle generalizes the maximum entropy principle above. Here, one is given a reference density $q(x)$ and wishes to determine the density $p(x)$ (not presumed to be an exponential family model) so as to minimize the Kullback-Leibler divergence[2]

$$D(p||q) = \int p(x) \log \frac{p(x)}{q(x)} dx \tag{5}$$

again subject to a set of moment constraints of the form $E_p t(x) = \eta^*$. Again, by the method of Lagrange multipliers, the optimum density is an exponential family density $p^*(x) = f(x;\theta^*)$ with statistics $t(x)$ but now having base measure $q(x)$. The Lagrange multipliers $\theta^*$ are again determined by solution of the moment-matching problem $\eta(\theta) = \eta^*$.

A dual version of the above KL-minimization problem may be stated as follows. Suppose now that we are given some reference density $p(x)$ and now wish to determine the member $\theta$ of a specified exponential family of models which minimizes the KL-divergence $D(p||f(\cdot;\theta))$ over the set of admissible $\theta$ for that family. That is we wish to select the best approximation of $p$ within the given exponential family specified by some measure $q(x)$ and statistics $t(x)$. It may be shown that this reduces to the moment-matching problem in the sense that the optimum $\theta^*$ then satisfies $\eta(\theta^*) = E_p t(x)$. So, if we can evaluate the reference moments $\eta^* = E_p t(x)$ then the "projection" of $p$ to the exponential family has moment coordinates $\eta^*$. Determining the corresponding exponential coordinates then reduces to the moment matching problem.

[1] Otherwise, a reduced set of statistics may be employed which are minimal.

[2] Kullback-Leibler divergence is sometimes referred to as relative or cross entropy as it may be considered as an invariant form of entropy.

*Maximum Likelihood.* This latter type of "KL-projection" is important in practice as it arises in the context of maximum-likelihood parameter estimation. Here, we observe some unknown density $p(x)$ through a set of independent identically distributed samples paths $x^{(1)}, \ldots, x^{(N)} \sim p$ and wish to determine the member of a given exponential family which maximizes the joint log-likelihood of the data.

$$\hat{\theta}_{ML} = \arg\max_{\theta} \sum_{k=1}^{N} \log f(x^{(k)}; \theta) \tag{6}$$

This likelihood maximization is equivalent to minimizing the KL-divergence $D(\tilde{p}||f(\cdot;\theta))$ where the empirical distribution $\tilde{p}(x)$ may be taken to be any density having the same moments as the data such that $E_{\tilde{p}}t(x) = \tilde{\eta}$ where the empirical moments $\tilde{\eta}$ are the sample-averaged statistics.

$$\tilde{\eta} = \frac{1}{N} \sum_{k=1}^{N} t(x^{(k)}) \tag{7}$$

The maximum-likelihood parameters then also minimize the KL-divergence from the empirical distribution and may be obtained by solving the moment-matching problems $\eta(\theta) = \tilde{\eta}$.

The reader is also referred to Amari [Ama01] and Csiszar [Csi75] for further material on information geometry and the related KL-projection problem.

### 2.2 Graphical Models

This section briefly introduces the notion of a graphical model [Lau96, Jor99] for a Markov random field and illustrates the connection to exponential family 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. Consider the vertices of this graph $i = 1, \ldots, |V|$ as indexing the elements of a random vector x such that random variable $x_i$ is the state at vertex $i$. The random process x is then said to be *Markov* with respect to $\mathcal{G}$ if it satisfies a certain set of conditional independence relations specified by the edge structure of the graph. These conditional independence relations may be specified in a number of equivalent forms. One form states is specified with respect to the neighborhood system defined by the graph. Let us say that two vertices are *neighbors* if they are linked to some edge $\langle ij \rangle \in \mathcal{E}$. The Markov property holds if the state of each vertex is conditionally independent of all other non-neighboring states given the state of just its neighbors. In this manner, the edges of the graph then represent direct interactions between random variables.

The Hammersley-Clifford theorem then establishes an important equivalence between this Markov structure of the field and a certain factorization property of the probability density function $p(x)$. Let us say that the pdf $p(x)$ factors according to $\mathcal{G}$ if it may be expressed as a product of positive "potential" functions each depending upon the states of some fully-connected subset of vertices. If we let $\mathcal{C}$ denote the set of such "cliques" of $\mathcal{G}$ and $\psi_C(x_C)$ denote the potential function on clique $c \in \mathcal{C}$ then we may express the factorization by

$$p(x) = \frac{1}{Z(\psi)} \prod_{c \in \mathcal{C}} \psi_c(x_c) \tag{8}$$

where $Z(\psi)$ is just a normalization constant. Without any loss of generality, the set $\mathcal{C}$ may be restricted to just those maximal cliques which are not a subset of any other clique so as to reduce the required number of potential functions. Hammersley-Clifford says that x is Markov on $\mathcal{G}$ if and only if it's pdf factors as above. Thus, the Markov structure of x allows for more compact

representation of the pdf $p(x)$ through a set of local influence functions specifying the relative compatibility of local state configurations of sets of neighboring vertices.

This suggests the following family of "exponential" graphical models. If we restrict the statistics $t(x)$ to consist solely of "local" statistics on cliques of vertices $t_c(x_c)$ then the exponential family pdf given earlier factors as above with potential functions

$$\psi_c(x_c) = \exp\{\theta_c \cdot t_c(x_c)\} \tag{9}$$

and normalization constant

$$Z(\psi) = \exp\{\varphi(\theta)\}. \tag{10}$$

Such an exponential family model is then Markov with respect to $\mathcal{G}$. Conversely, any Markov random field can be described as such an exponential family model provided it can be naturally parameterized in such a way that the log potentials vary linearly in the parameters. For instance, models of this type often arise in statistical mechanics through the use of the Gibbs distribution where the potentials then correspond to actual physical energy functions describing forces upon/between the states of the system. It is this especially simple type of parameterization which proves especially tractable in the information geometry of exponential families.

### 2.3 Gaussian Processes

In this section we briefly describe the Gaussian process as an exponential family model and introduce the so-called information filter representation. This representation provides a convenient graphical model for Gauss-Markov random fields.

Consider a Gaussian random vector $\mathrm{x} \sim \mathcal{N}(\mu, \Sigma)$ with mean vector $\mu = E\{\mathrm{x}\}$ and covariance matrix $\Sigma = E\{\mathrm{x}\mathrm{x}'\} - \mu\mu'$. Provided the covariance matrix is positive definite, we may express this as a regular exponential family model with statistics $t(x) = (x, xx')$ as follows. It proves convenient to reparameterize the Gaussian density in the so-called *information filter* form $\mathrm{x} \sim \mathcal{N}^{-1}(h, J)$ which is related to the mean-covariance form as shown below.

$$h = \Sigma^{-1}\mu \tag{11}$$

$$J = \Sigma^{-1} \tag{12}$$

The density function within this parameterization is expressed as

$$p(x) = \exp\{-\frac{1}{2}x'Jx + h'x - \varphi(h, J)\} \tag{13}$$

where

$$\varphi(h, J) = -\frac{1}{2}\{h'J^{-1}h + \log|J| + n\log 2\pi\}. \tag{14}$$

This corresponds to a regular exponential family with statistics $t(x) = (x, xx')$ and exponential parameters $\theta = (h, -J/2)$. The cumulant function is then given by $\varphi(\theta) = \varphi(h, J)$ above. Hence, the information filter form of the Gaussian density essentially corresponds to the exponential parameters $\theta$ while the moment parameters are then trivially related to the mean-covariance form of the Gaussian density by $\eta = (\mu, \Sigma + \mu\mu')$. Note that the statistics of this formulation of the Gaussian process as an exponential family model are entirely comprised of linear and quadratic "singleton" affects $t_i(x) = (x_i, x_i^2)$ and "pairwise" interactions $t_{i,j}(x_i, x_j) = x_i x_j$.

Now consider the affect of imposing constraints upon the Markov structure of a Gaussian process. In keeping with the equivalence of the "factor" structure of an exponential family density

and the corresponding "Markov" structure of the process, we have that two variables $x_i$ and $x_j$ of a Gaussian process are conditionally independent if and only if there are no pairwise statistics linking those two variables. This requires that the corresponding exponential parameter $\theta_{ij} = -J_{ij}$ be forced to zero. This point is made explicit for Gaussian processes by the following considerations. The partial correlation coefficient between variables $x_i$ and $x_j$ is defined as the usual correlation coefficient of the conditional density between those variables given the state $\mathrm{x}_{ij}^c$ of the remaining variables.

$$\rho(\mathrm{x}_i, \mathrm{x}_j | \mathrm{x}_{ij}^c) = \frac{\mathrm{cov}(\mathrm{x}_i, \mathrm{x}_j | \mathrm{x}_{ij}^c)}{\sqrt{\mathrm{cov}(\mathrm{x}_i | \mathrm{x}_{ij}^c)\mathrm{cov}(\mathrm{x}_j | \mathrm{x}_{ij}^c)}} \tag{15}$$

For Gaussian processes this coefficient is related to the conditional mutual information below.

$$I(\mathrm{x}_i; \mathrm{x}_j | \mathrm{x}_{ij}^c) = -\frac{1}{2}\log(1 - \rho^2(\mathrm{x}_i, \mathrm{x}_j | \mathrm{x}_{ij}^c)) \tag{16}$$

Consequently, the two variables $x_i$ and $x_j$ are conditionally independent given the state of the remaining variables $\mathrm{x}_{ij}^c$ if and only if the partial correlation coefficient (and corresponding conditional mutual information) are zero. However, the partial correlation coefficient is readily evaluated directly from the relevant entries of the information matrix

$$\rho(\mathrm{x}_i, \mathrm{x}_j | \mathrm{x}_{ij}^c) = -\frac{J_{ij}}{\sqrt{J_{ii}J_{jj}}} \tag{17}$$

and is hence zero if and only the corresponding off-diagonal entry of $J$ is zero. The essential point here is that the sparsity structure of the inverse covariance matrix $J$ then reflects the Markov structure of the process. Hence, the information filter form $(h, J)$ of a Gauss-Markov process provides a natural compact graphical model for the process with $J$ a sparse matrix. Note also that, due to the quadratic form of the log-density, only "pairwise" potentials are required to characterize Gauss-Markov processes regardless of the presence or absence of higher-order cliques.

The description of a Gauss-Markov process as an exponential model with respect to the graphical structure $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is illustrated below. The statistics may be chosen as below.

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

The parameters are related to the information filter form $(h, J)$ as below.

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

The moments are related to the standard Gaussian moment parameterization.

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

Inference of the moments $\eta(\theta)$ of model $\theta$ 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.

## 3 Iterative Methods for M-Projection

We now consider iterative algorithms for solving the moment-matching problem which, from the information geometry viewpoint, may also be considered as an M-projection algorithm. This allows us to restate the moment-matching problem (solving the nonlinear system $\eta(\theta) = \eta^*$) as the following convex programming problem.

Consider an exponential family $\mathcal{F} = \{f(\cdot;\theta)|\theta \in \Theta\}$ specified with respect to given base measure $b(x)$ and set of statistics $t(x)$ with $\Theta$ denoting the set of admissible parameters $\theta$ and $\eta(\Theta)$ denoting the set of achievable moments. Given $\eta^* \in \eta(\Theta)$, consider the following nonlinear program.

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

Here, $D(\eta^*||\theta)$ abbreviates the KL-divergence between the (unknown) density $f^* \in F$ having moment coordinates $\eta^*$ and the density $f(\cdot;\theta)$. For an exponential family, this KL-divergence may be expressed as

$$D(\eta^*||\theta) = \varphi^*(\eta^*) + \varphi(\theta) - \theta^* \cdot \eta \tag{18}$$

where $\varphi(\theta)$ is the cumulant function defined previously and $\varphi^*(\eta)$ is its convex conjugate which equals minus the entropy as a function of the moments. As these are both convex functions, this shows that $D(\eta^*||\theta)$ is convex in either $\eta^*$ or $\theta$. This convexity is strict under the assumption of minimality. Furthermore, the set $\Theta$ of admissible moment parameters is convex. Hence, the program (P) is a convex programming problem. Furthermore, the KL-divergence $D(p||q)$ is non-negative and is zero if and only if $p(x) = q(x)$ for essentially all $x$ (except on a set of measure zero). Hence, the value of (P) is zero and the unique global minimizer $\theta^*$ must satisfy $f^*(\cdot) = f(\cdot;\theta^*)$ which (for a minimal representation) occurs if and only $\eta(\theta^*) = \eta^*$. Hence, solving the convex programming problem is equivalent to solving the moment matching problem. Of course, the term $\varphi^*(\eta^*)$ is constant and may be omitted without changing the location of the minimum. The objective function is then $g(\theta) = \varphi(\theta) - \eta^* \cdot \theta$ and its minimum value is the entropy $h(\eta^*) = -\varphi^*(\eta^*)$.

Several other properties of the programming problem are noteworthy. First, due to the strict convexity of the objective function any stationary point (zero of the gradient) is a global minimum. Hence, we may actually implement strictly gradient-based minimization algorithms which do not require evaluation of the objective function (which would entail evaluation the potentially intractable cumulant function). Second, consider the feasible set of admissible parameters $\theta \in \Theta$ which is defined by the constraint that $\varphi(\theta) < \infty$. Observe that since the KL-divergence includes $\varphi(\theta)$ it also tends towards infinity as $\theta$ approaches any boundary point of $\Theta$. Hence, this normalizable constraint is automatically enforced within this formulation of the moment-matching problem (i.e., KL-divergence provides a natural "barrier" function for this feasible set).

We now note certain useful differential relations concerning the cumulant function which are easily derived from the definition. The gradient of the cumulant function generates the moments.

$$\nabla_\theta \varphi(\theta) = \eta(\theta) \tag{19}$$

The Hessian of the cumulant function generates the Fisher information matrix defined as the covariance of the 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 the KL-divergence is given by the difference in the moments.

$$\nabla_\theta D(\eta^*||\theta) = \eta(\theta) - \eta^* \tag{20}$$

The Hessian of the KL-divergence is also the Fisher information.

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

Hence, if we are able to implement the moment calculation $\eta(\theta)$ in a tractable manner for a given exponential family, then we are prepared to implement gradient-based minimization method for minimizing $D(\eta^*||\theta)$ thus solving the moment-matching problem and are assured of the convergence of the method due to the strict convexity of the objective function in $\theta$. Furthermore, if we are able to also evaluate higher order moments of the statistics required by the Fisher information calculation, then we may take advantage of this ability to implement Newton's method so as to obtain more rapid convergence near the minimum (but at the expense of inverting the Fisher information).

For Gaussian process, evaluation of the Fisher information matrix is simplified due to the fact that higher order moments between jointly Gaussian random variables always reduce to expressions involving just first and second order moments. In particular, the third order moments of the zero-mean Gaussian process $\tilde{\mathrm{x}} \equiv \mathrm{x} - \mu$ are all zero

$$E\{\tilde{\mathrm{x}}_i\tilde{\mathrm{x}}_j\tilde{\mathrm{x}}_k\} = 0 \tag{21}$$

while the fourth order moments are given by

$$E\{\tilde{\mathrm{x}}_i\tilde{\mathrm{x}}_j\tilde{\mathrm{x}}_k\tilde{\mathrm{x}}_l\} = \Sigma_{ij}\Sigma_{kl} + \Sigma_{ik}\Sigma_{jl} + \Sigma_{il}\Sigma_{jk} \tag{22}$$

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

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

The latter formulas for higher order moments are only evaluated as needed, e.g. $G_{\langle ij\rangle;k}, G_{ii;k}, G_{\langle ij\rangle;\langle kl\rangle}, G_{\langle ij\rangle;kk}$, and $G_{ii;\langle kl\rangle}$

We now briefly describe the (standard) nonlinear programming techniques we have applied to the nonlinear program (P). We perform minimization of $\varphi(\theta) - \eta^* \cdot \theta$ employing earlier gradient $g(\theta)$ and Hessian $G(\theta)$ evaluators and the following standard[3] methods.

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

[3]Bertsekas, 95.

- *Conjugate Gradients.* uses "non-jamming" direction update and performs conjugacy test for early "restarts" with threshold 0.05.
- *Preconditioned Conjugate Gradients.* as above with preconditioning matrix $M$ chosen as either the inverse diagonal $M = Diag(G(\theta))^{-1}$ or as "full" inverse $M = G(\theta)^{-1}$.
- *Newton's Method.* without line-minimization.

All methods are initialized by m-projection of $\eta^*$ to "fully factorized" (disconnected) family.

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

This is just the product of the marginals under the given moments.

We also implement the standard method of Iterative Proportional Fitting (IPF) described in Csiszár but do not describe the algorithm in here. The main idea is that IPF performs coordinate descent on set of exponential parameters associated with a given edge and may be interpreted as performing projections onto intersecting manifolds corresponding to subsets of the moments constraints.

### 3.1 Application to ML-Estimation of Gauss-Markov Processes

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

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

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 $(h, J)$.

We implemented the optimization using six main techniques: gradient descent (GD), conjugate gradient (CG), iterated proportional fitting (IPF), preconditioned conjugate gradient using the diagonal of the Hessian (Diag-PCG) and the full Hessian (PCG-Full), and Newton's method (Newton). All except IPF are unconstrained descent techniques with parameter updates given as:

$$\theta^{k+1} = \theta^k + \alpha^k d^k$$

For GD, $d^k = -\nabla f(\theta^k)$ and $\alpha^k$ is determined by line minimization. CG determines $d^k$ by taking the gradient vector and making it Q-conjugate to previous search directions. $\alpha^k$ is determined by line minimization. PCG methods are just CG methods with a preconditioning matrix S applied to $\theta$ to reduce the condition number of its Hessian to help with convergence. Diag-PCG uses only the diagonal elements of the Hessian while PCG-Full uses the full Hessian. Newton is similar to PCG-Full. $d^k = -\nabla^2 f(\theta^k)^{-1} \nabla f(\theta^k)$ and we don't do line minimization ($\alpha^k = 1$).

We tested on a variety of graph structures to evaluate robustness. We used nxn grids which are common in image processing (Figure 1). These have lots of local loops but interactions terminate

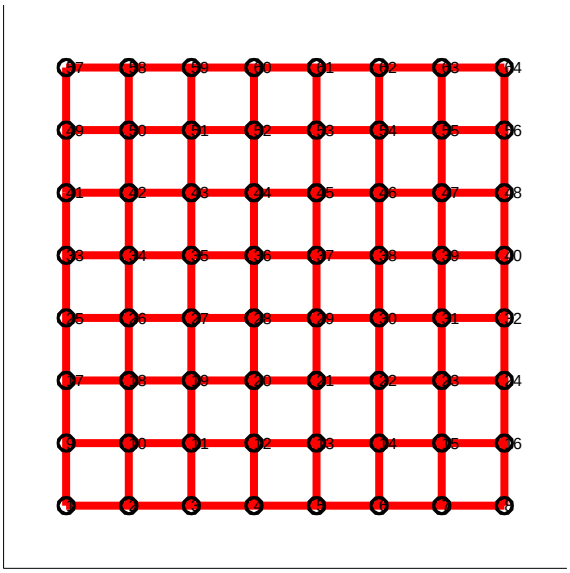

Figure 1: 8x8 Grid Interaction Graph.

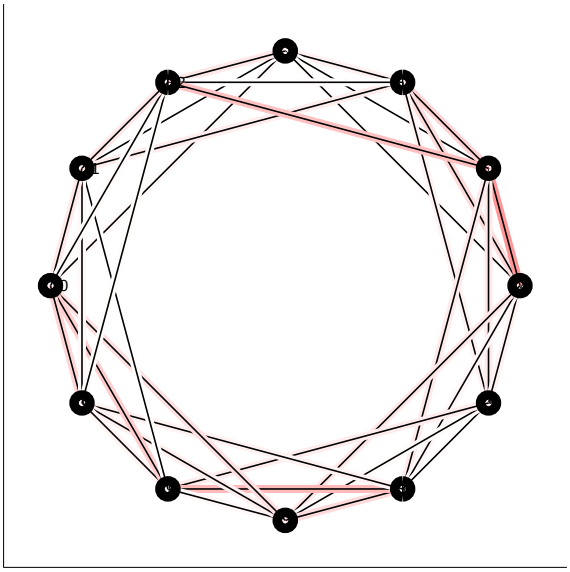

Figure 2: 12-Node 3rd-Order Loop Interaction Graph.

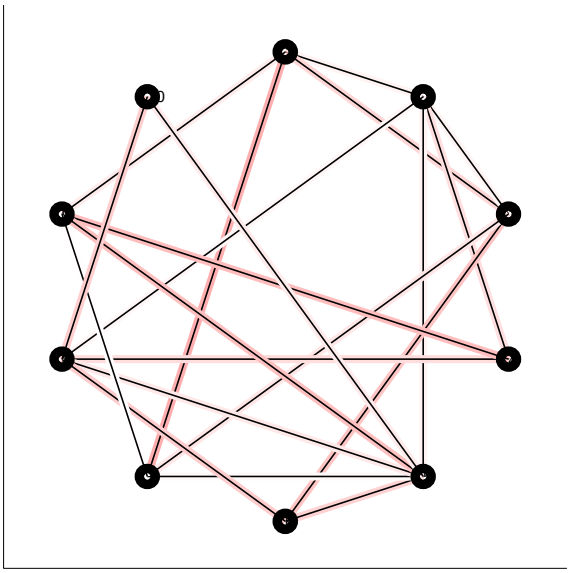

Figure 3: Sparse 10-Node 11-Edge Random Interaction Graph.

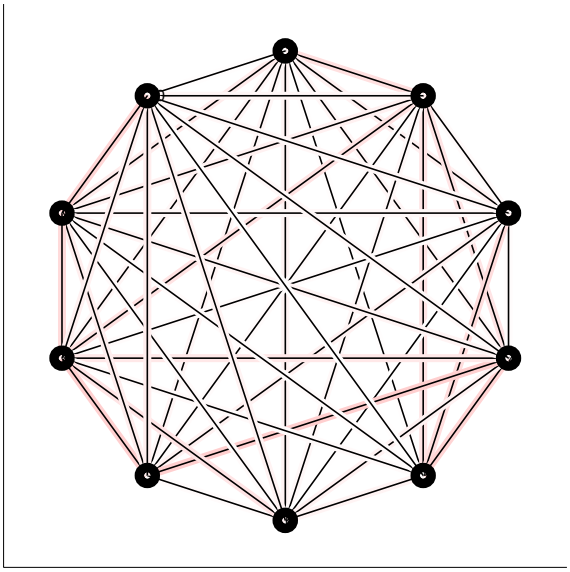

Figure 4: Dense 10-Node 35-Edge Random Interaction Graph.

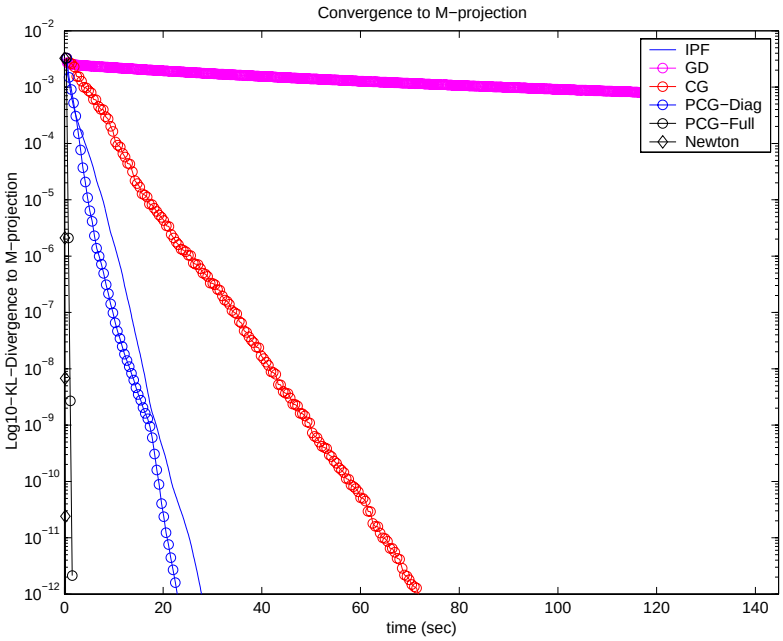


Figure 5: 8x8 Grid Example. log10-KL-Error vs. Run-Time.

at the boundaries. We also used n-node k-loops. Each node is connected to nodes within k of it (Figure 2). Graphs with random edges were tested to ensure that certain techniques were not put at an advantage due to special structure in the graph. We tested using both sparse (Figure 3) and dense (Figure 4) graphs.

In comparing the various optimization techniques, we found differing behavior based on graph structure, but a few trends were evident. Using the full Hessian information can lead to vast improvements in speed, but even just using the diagonal elements as a scaling matrix can produce significant improvements. We can fairly conclusively declare Newton and PCG-Full as the winners in the examples that we used. GD almost always performs the worst. CG and IPF tend to occupy the next worst slots, though IPF occasionally manages to surpass even Diag-PCG.

IPF performs nearly identically to Diag-PCG on the 8x8 grid but sits between GD and CG on the k-loop (and is as slow as GD at the beginning). Because IPF is essentially a coordinate descent algorithm, it takes time for information from one node to propagate to the rest of the nodes. With the longer, more global loops present in the k-loop, IPF needs more time to converge due to the more global interactions.

In Figures 7 and 8 we observe the same behavior in IPF as noted before: IPF has trouble with global interactions. In Figure 9, we see that IPF is also terrible with dense graphs. It is unclear whether the problems that IPF experiences are solely from the global interactions (many of which are present in a dense graph), or whether the sheer number of interactions between nodes exacerbates things. Of course, most real world applications will be much sparser than our dense examples since the whole point of graphical models is to build sparse models. One thing that we observed is that even though Newton consistently beats PCG-Full in convergence time, PCG-Full consistently beats Newton in number of iterations to convergence. We conclude that the line search is taking up a substantial amount of processing time. It would be interesting to see how PCG-Full and the other line search algorithms perform with inexact line search techniques.

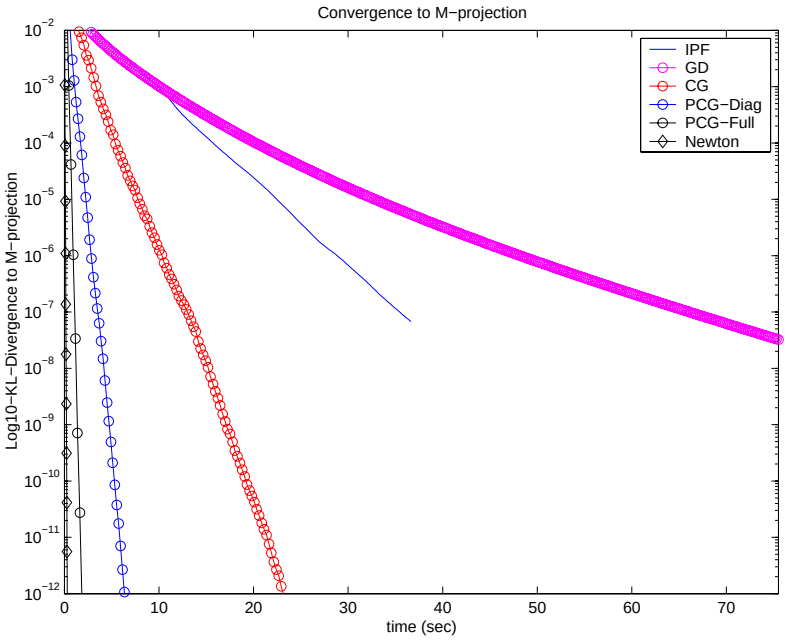


Figure 6: 40-Node 3rd-Order Loop Example. log10-KL-Error vs. Run-Time.

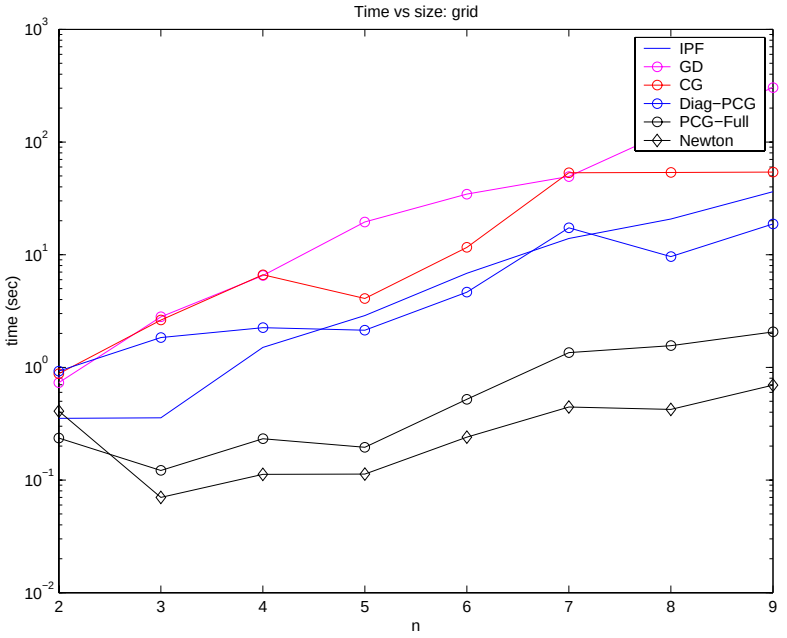


Figure 7: nxn grid, convergence time vs n.

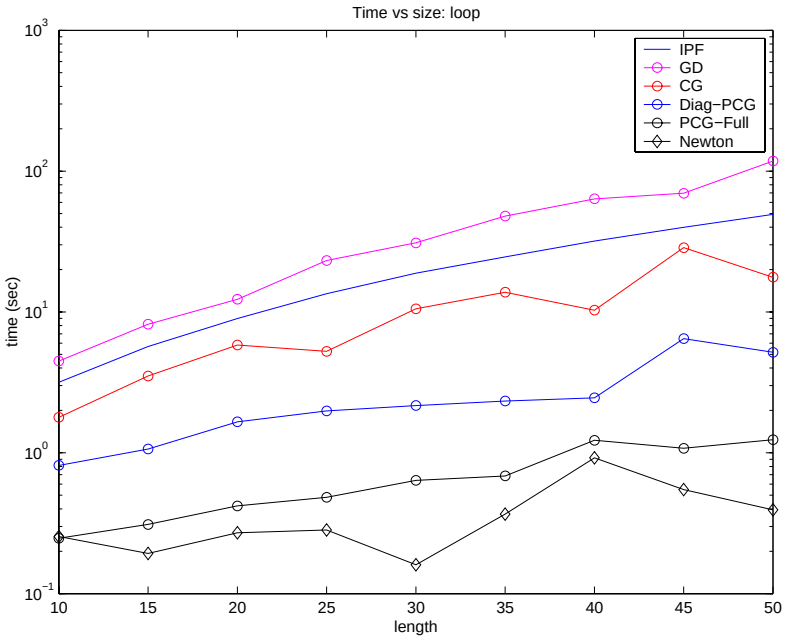


Figure 8: n-node loop, 3rd order, convergence time vs n.

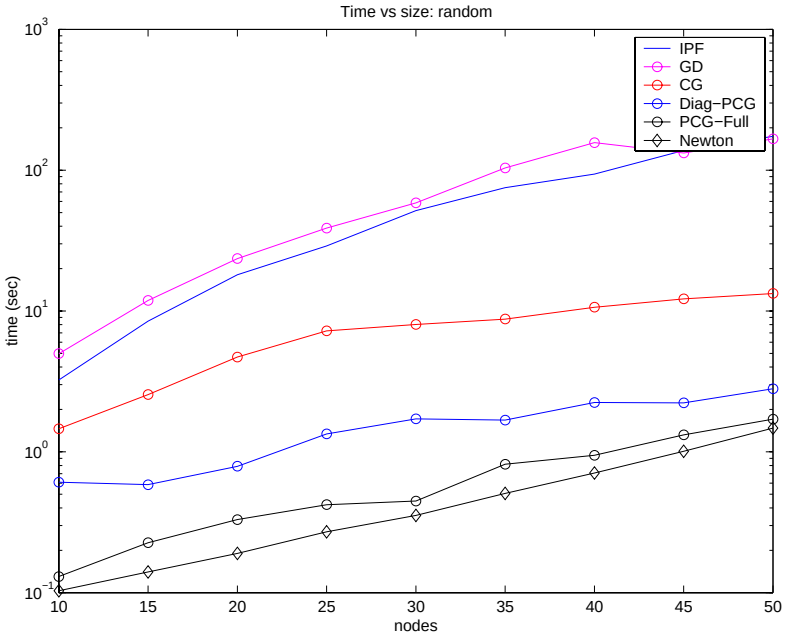


Figure 9: n-node random graph, dense, convergence time vs n.

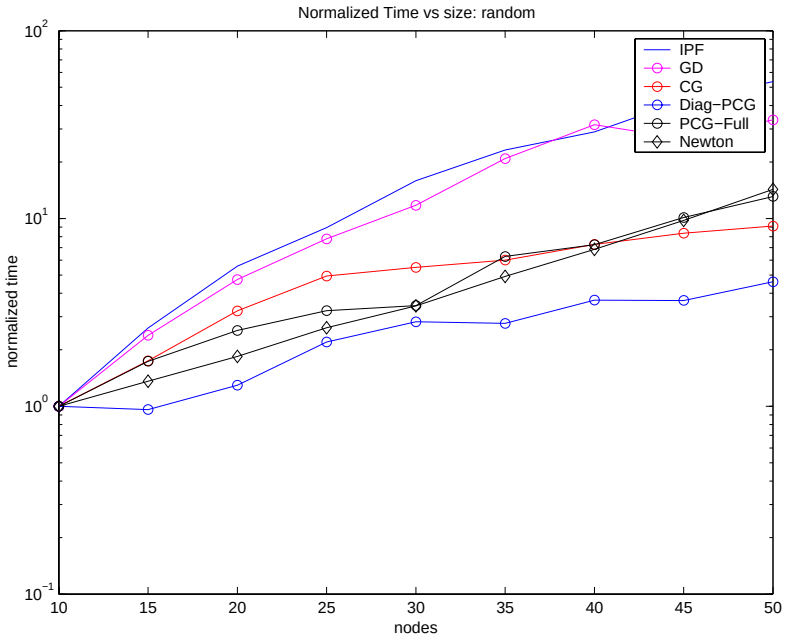


Figure 10: n-node random graph, dense, normalized convergence vs n.

As problem size scales, it may be that other techniques that don't require inversion of a large matrix may catch up in speed. Specifically we would expect Newton and PCG-Full to lose in competitiveness as problem size scales. Some evidence for this can be seen in Figure 10 which depicts normalized scaling for dense random graphs. We see that Diag-PCG scales much better than methods that use the full Hessian, and at the end, even CG pulls ahead of the methods that use the full Hessian. For the most part though, the scaling we would expect is not really shown in our data. This is because the computation of the gradient vector requires inversion of the J matrix which also scales with the size of the problem. It may be that approximate inference techniques will allow faster solutions without a substantial decrease in accuracy and allow the other algorithms to scale better.

## 4 M-Projections for Structure Estimation

Structure estimation belongs to the wide field of system identification, which deals with forming models to describe real phenomena based on observed data. The importance need not be emphasized, since this is one of the main goals of all natural sciences. However, the philosophy is somewhat different - the task is not to find the true model for a given natural system (an impossible problem in general), but to find a model which is adequate to describe and predict the set of variables of interest to the practitioner. Thus a model is never truly "correct", it can only be a good fit to the observed data. Two pathways for system identification are "black-box" and physical insight. In the latter, the variables and their relations have physical meaning, whereas in the black-box approach, some parameters may have no interpretation whatsoever, and serve the sole purpose of giving an extra handle to get a good fit to the data. In this section, we are mainly be concerned with this latter view of the field.

The procedure for model identification includes the following steps: collecting data, postulating

a set of candidate models, and the criterion of goodness of fit of the model to the data, and a procedure to find the model in the set that best fits the observed data. For structure estimation, the set of candidate models is divided into families of models with the same structure, and the task is not to get an individual model, but rather a structural family that best fits the data. Of course, the problem is intimately related to system identification, and the subdivision of the set into families according to structure is quite arbitrary, but in the context of graphical models it turns out to be very meaningful. The family of models that we will deal with in this work are the Gauss-Markov random fields. The question whether or not this family of models is appropriate for a particular application will not be treated here, however there are many applications for which GMRF's are a natural framework, for example multiscale methods for linear estimation, image processing, oceanography, meteorology, geophysics, control, time series analysis, communications, and coding, to name a few.

The data that we obtain consists of $N$ samples of the process, where each sample consists of $M$ variables. The task is to determine the Markov structure of the statistical relationship among the variables, for example to conjecture a general sparse graph as a structure, or to choose a grid, a chain, a tree or a loop to describe the data. A widely used measure of fidelity of a model to the data is the Kullback-Leibler divergence discussed earlier. If the structure of the model was known and the number of unknown parameters fixed, then the KL-divergence framework would produce the Maximum Likelihood (ML) model, and would be very appropriate. This is the approach considered in the previous section.

However, note that among the family of models with different structures the model that has a more complicated structure is preferred to a simpler model. Any sparse model is embedded in the full graph model, which has more parameters, and thus all the sparse structures belong to the full structure (with some interactions set to zero). In the Gaussian case, the Maximum Likelihood model from the data is determined by the sample mean and covariance, and will typically have a full structure. Even if the data was generated by a chain model, due to its finite sample size the elements of the inverse covariance matrix which, due to conditional independence, should be zero will not be, and the structure is lost. This phenomenon comes under the umbrella of overfitting noisy data. However, a way around this involves the use of the Akaike Information Criterion (AIC) [Aka73], or some related criteria such as the Bayesian Information Criterion (BIC) [Sch78] or the Minimum Description Length (MDL) principle. All of these criteria balance the contrary goals of maximizing data fidelity while minimizing model complexity.

We focus on the simpler AIC and BIC methods, each of which measure data fidelity by the log-likelihood and model complexity by the number of model parameters. We may express the data fidelity objective as the equivalent goal of minimizing the KL-divergence $D(\tilde{\eta}_{\mathcal{F}}||\theta_G)$ where $\tilde{\eta}_{\mathcal{F}}$ are the empirical moments presuming a "full" interaction graph $\mathcal{F}$ and $(\mathcal{G}, \theta_{\mathcal{G}})$ is a hypothesized sparse model. For Gaussian processes, the empirical moments correspond to the sample mean and covariance $(\tilde{\mu}, \tilde{\Sigma})$. For the information filter form of Gauss-Markov random fields with Markov structure $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ the number of model parameters is twice the number of vertices (variables) plus the number edges (interactions).

$$K_{\mathcal{G}} \equiv 2|\mathcal{V}| + |\mathcal{E}| \tag{23}$$

Both the AIC and BIC may then be fit into the common framework of

$$\text{minimize} \quad C(\theta_{\mathcal{G}}) \equiv D(\tilde{\eta}_{\mathcal{F}}||\theta_{\mathcal{G}}) + \delta K_{\mathcal{G}} \tag{24}$$

$$\text{w.r.t} \qquad (\mathcal{G}, \theta_G) \tag{25}$$

where the parameter $\delta \geq 0$ is chosen as either

$$\delta_{AIC} = \frac{1}{N} \tag{26}$$

or

$$\delta_{BIC} = \frac{\log N}{2N} \tag{27}$$

according to the favored criterion. Note that the $\delta$ parameter expresses the degree to which we emphasize the goal of minimizing model complexity relative to the goal of maximizing the fidelity of the model to the data so that large values of $\delta$ favors sparser models. Both methods have $\delta$ vanishing as the number of samples becomes large. Hence, the more data we have for model identification the more apt we are to allow complex models. The BIC decreases $\delta$ more slowly then the AIC and hence favors more complex models.

In either case, a direct approach for determining the best model apparently results in a very expensive combinatorial search among all possible permutations of models with removed edges, which for data with large number of variables is impractical. We propose a greedy search procedure to find a "good" model which is tractable.

For a given structure $\mathcal{G}$ the complexity $K_{\mathcal{G}}$ is fixed such that the best parameters $\theta_{\mathcal{G}}$ within this structure is given by $\theta^*_{\mathcal{G}}$ minimizing the KL-divergence $D(\tilde{\eta}_F||\theta_{\mathcal{G}})$ which, as described in previous sections of the paper, can be accomplished by an M-projection. Thus the naive intractable approach consists of computing the M-projection of $\tilde{\eta}_{\mathcal{F}}$ to every lower order structure $\mathcal{G}$ and selecting the $\delta$-optimal structure $\mathcal{G}^*$ minimizing $D(\tilde{\eta}_{\mathcal{F}}||\theta^*_{\mathcal{G}}) + K_G$.

We propose a very different procedure, which relies upon an inductive decomposition of the cost $C$ and also exploits a lower-bound of the KL-divergence to a model in the rank 1 lower submanifold which holds in the Gaussian case.

The *Pythagorean theorem* of information geometry states that for embedded graphs $\mathcal{G}_1 \subset \mathcal{G}_2$ the KL-divergence decomposes as

$$D(\theta_{\mathcal{G}_1}||\theta_{\mathcal{G}_2}) = D(\theta_{\mathcal{G}_1}||\theta^*_{\mathcal{G}_2}) + D(\theta^*_{\mathcal{G}_2}||\theta_{\mathcal{G}_2}), \tag{28}$$

where $\theta^*_{\mathcal{G}_2}$ is the m-projection of $\theta_{\mathcal{G}_1}$ onto $\mathcal{G}_2$. This allows the model selection metric to be evaluated inductively from the full model by projections onto submanifolds with 1 less rank. Consider a sequence of embedded graphs $\mathcal{G}_K \subset \mathcal{G}_{K-1} \subset \ldots \mathcal{G}_1 \subset \mathcal{G}_0 \equiv \mathcal{F}$ where $\mathcal{G}_k$ has removed $k$ edges from $\mathcal{F}$. The cost function then decomposes as

$$C(\theta^*_{\mathcal{G}_k}) = K_{\mathcal{F}} + \sum_{k=0}^{K-1} \Delta(\theta^*_{\mathcal{G}_k}||\theta^*_{\mathcal{G}_{k+1}}) \tag{29}$$

where

$$\Delta(\theta^*_{\mathcal{G}_k}||\theta^*_{\mathcal{G}_{k+1}}) = D(\theta^*_{\mathcal{G}_k}||\theta^*_{\mathcal{G}_{k+1}}) - \delta \tag{30}$$

Hence, so long as there exists a favorable edge removal such that $D(\theta^*_{\mathcal{G}_k}||\theta^*_{\mathcal{G}_{k+1}}) < \delta$ we may decrease the cost function by projecting to a lower order model. Also, the most favorable edge removal is the one which minimizes the KL-divergence $D(\theta^*_{\mathcal{G}_k}||\theta^*_{\mathcal{G}_{k+1}})$.

This decomposition suggests that we proceed in the following fashion: First, set $\mathcal{G}$ to be the full graph $\mathcal{F}$ and determine $\theta^*_{\mathcal{G}}$ as the ML-estimate given by,

$$\tilde{h} = \tilde{\Sigma}^{-1}\tilde{\mu} \tag{31}$$

$$\tilde{J} = \tilde{\Sigma}^{-1} \tag{32}$$

Second, consider all potential edge removals $\mathcal{G} \setminus \langle ij \rangle$, the rank-1 lower embedded submanifold, and select the edge $\langle ij \rangle^*$ which minimizes $D(\theta^*_{\mathcal{G}} || \theta^*_{\mathcal{G} \setminus \langle ij \rangle})$. Note that this requires an m-projection for each edge in $\mathcal{G}$. Finally, if the incremental KL-divergence of the best edge-removal is less than $\delta$, then replace $(\mathcal{G}, \theta^*_{\mathcal{G}})$ by $(\mathcal{G} \setminus \langle ij \rangle^*, \theta^*_{\mathcal{G} \setminus \langle ij \rangle^*})$ and search for another favorable edge removal. Otherwise, the cost is increasing by going to the lower order models so we terminate with the current estimate. Note that $\delta$ effectively sets a threshold on the incremental KL-divergence controlling whether or not to continue order reduction.

This approach already reduces the complexity of the procedure dramatically, but an expensive search for the best structure in the 1-lower submanifold still remains. The following lower bound for the KL-divergence can be exploited to get even lower complexity:

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

Here, $I_{\langle ij \rangle}$ is the mutual information between vertices $i$ and $j$ conditioned on the rest of the graph while $G \setminus < st >$ denotes the graph $G$ with edge $< s, t >$ removed. This relation holds for any member of the exponential family of models. For the Gaussian family, this mutual information is related to the partial correlation coefficient and hence may be calculated from the relevant entries of the information matrix $J$.

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

Thus, we only need to consider the edges which have low conditional information terms, which can be readily computed. In particular, when searching for favorable edge removals, we need only consider those candidate edges having $I_{\langle ij \rangle} < \delta$. Furthermore, as we evaluate these candidate edge removals, if we should find a favorable removal with incremental divergence $d < \delta$, we may then eliminate any remaining untried candidates having $I_{\langle ij \rangle} > d$ as these cannot prove more favorable.

Combining the two ideas the following algorithm ensues given $(\tilde{\eta}_{\mathcal{F}}, \delta)$:

1. Start with the full model: set $(\mathcal{G}, \theta^*_{\mathcal{G}}) = (\mathcal{F}, \theta^*_{\mathcal{F}})$, where $\theta^*_{\mathcal{F}}$ is the unconstrained ML estimate from $\tilde{\eta}_{\mathcal{F}}$.
2. Evaluate $I_{\langle ij \rangle}$ for all $\langle ij \rangle$ in $\mathcal{G}$ and find the set $\Upsilon$ of candidates edges s.t. $I_{\langle ij \rangle} < \delta$.
3. Initialize $d^* = \delta$ and $\langle ij \rangle^* = \emptyset$.
4. While $\Upsilon$ in non-empty:
   (a) Select trial $\langle ij \rangle \in \Upsilon$ with minimum $I_{\langle ij \rangle}$.
   (b) compute $\theta^*_{\mathcal{G} \setminus \langle ij \rangle}$, the M-projection of $\theta^*_{\mathcal{G}}$ onto $G \setminus \langle ij \rangle$, and the resulting KL-divergence $d_{\langle ij \rangle} = D(\theta^*_{\mathcal{G}} || \theta^*_{\mathcal{G} \setminus \langle ij \rangle})$.
   (c) If $d_{\langle ij \rangle} < d^*$, set $\langle ij \rangle^* = \langle ij \rangle$, $d^* = d_{\langle ij \rangle}$ and $\Upsilon = \{ \langle ij \rangle \in \Upsilon : I_{\langle ij \rangle} < d^* \}$.
   (d) Remove $\langle ij \rangle$ from $\Upsilon$.
5. If $d^* < \delta$, replace $(\mathcal{G}, \theta^*_{\mathcal{G}})$ with $(\mathcal{G} \setminus \langle ij \rangle, \theta^*_{\mathcal{G} \setminus \langle ij \rangle^*})$ and return to step (2). Otherwise, exit with solution $(\mathcal{G}, \theta^*_{\mathcal{G}})$.

This concludes discussion of our structure estimation approach. The following examples illustrate the results of applying this technique to a 12 node loop with 3rd order nearest-neighbor connections and constant interactions. Figures 11, 12, 13 correspond to $N = 1000$ sample paths

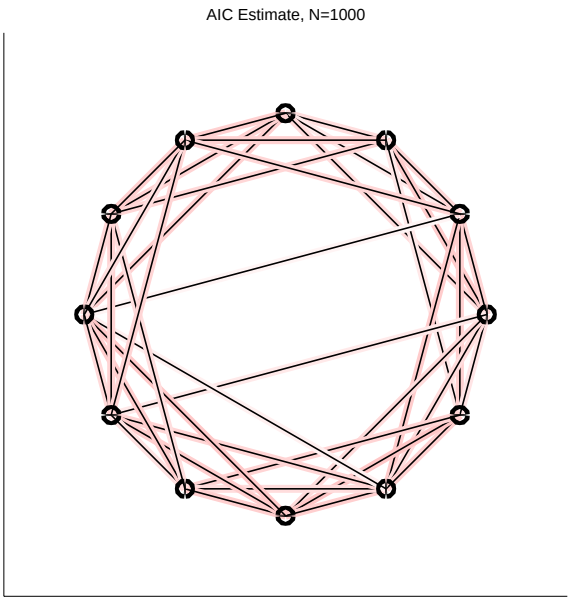


Figure 11: AIC Estimate for N=1000.

while Figures 14, 15, 16 correspond to $N = 10000$ sample paths. Figures 13 and 16 illustrate the AIC/BIC metrics as a function of the number of pruned edges and compares these metrics to the divergence of the estimate from the data and from the truth. Note that the BIC generates sparser estimates of graphical structure then the AIC. Also, the BIC has the property that it is an asymptotically optimal order estimator while the AIC is not. This is supported by our experiments. However, the intent of the AIC is to minimize on average the KL-divergence from the truth. The plots suggest that it is safer to overestimate the order of the model then to underestimate the order in this regard providing intuition as to why the AIC overestimates the order.

## 5 Conclusions

Our conclusions are summarized below:

- Moment-matching/M-Projection is a well-posed convex programming problem.
- Standard optimization techniques work quite well and are robust.
- Newton's method and the conjugate gradient methods typically outperform the standard IPF (coordinate descent) approach.
- Newton's method is most efficient for small graphs.
- Conjugate Gradients and Diagonal PCG may be more appropriate for larger problems provided efficient inference is available.
- The M-projection capability enables structure estimation with AIC/BIC.
- Our greedy edge-pruning approach for structure estimation seems to work well and is surprisingly efficient employing lower-bounds to eliminate edges from consideration.

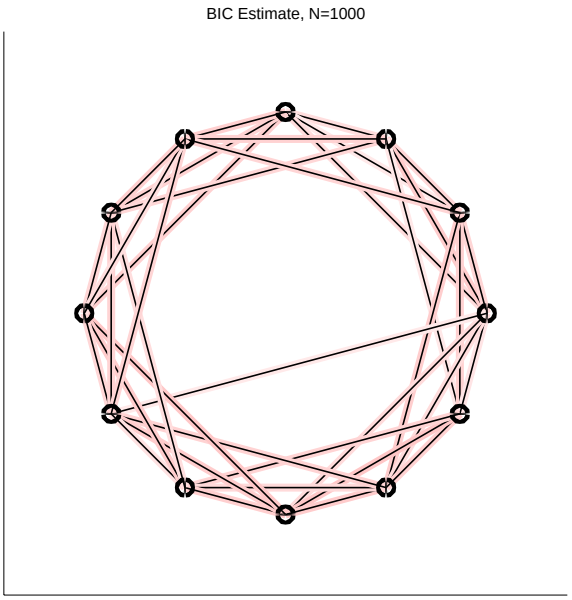


Figure 12: BIC Estimate for N=1000.

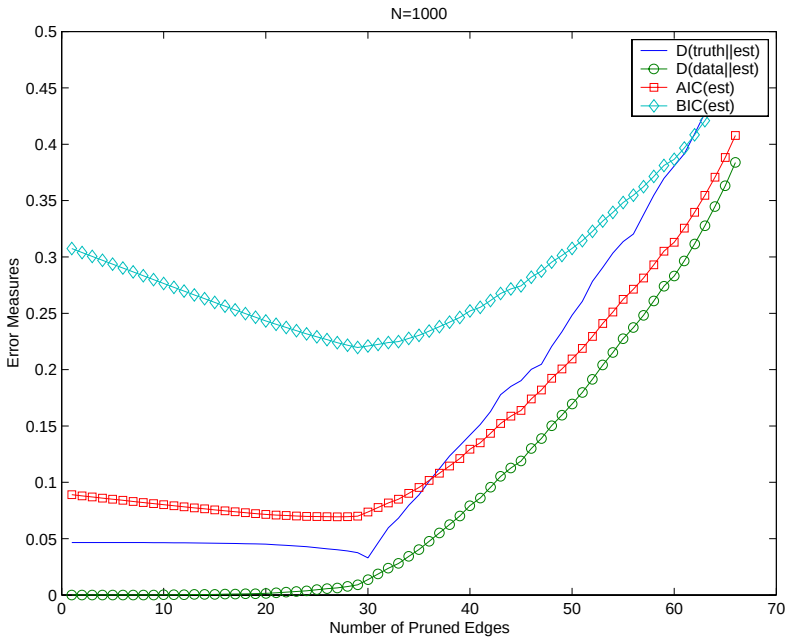


Figure 13: AIC/BIC Comparison for N=1000.

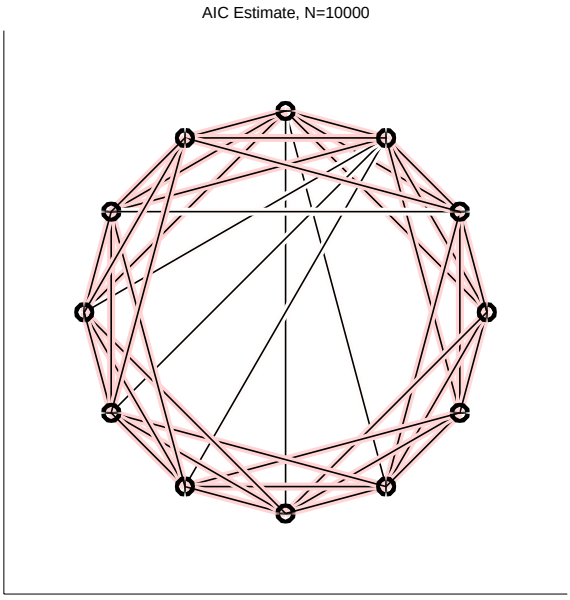


Figure 14: AIC Estimate for N=10000.

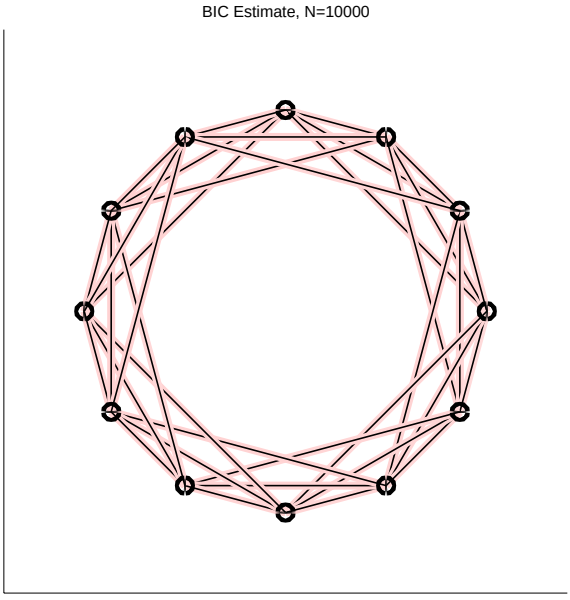


Figure 15: BIC Estimate for N=10000.

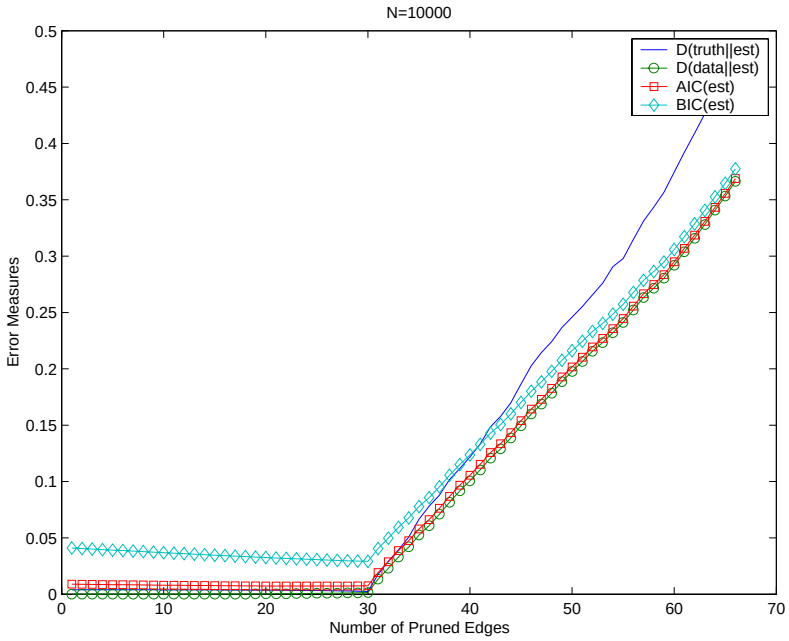


Figure 16: AIC/BIC Comparison for N=10000.

There are many potential directions for further research in this area. One area might focus on bridging the gap between PCG-Diag and the impressive performance of Newton's method by developing sparse approximations for the inverse Hessian. Also, a "dual" method of m-projection is to perform the projection within a higher-order family of exponential models (such as in the context of the structure estimation problem) containing the original model. The projection traces an "m-geodesic" containing the original model to the subfamily. This m-geodesic is the set of all models having the desired moments $\eta^*$ pertaining to the subfamily. An alternative to our approach is to determine the m-projection by tracing this m-geodesic to the subfamily. Finally, our "greedy" algorithm for structure estimation requires further analysis, but is presumably suboptimal. Nevertheless, the impressive performance of this approach suggests that more sophisticated dynamic programming techniques could achieve near-optimal structure estimation without requiring significantly more computation then the current method. Also, in addition the the "thinning" approach we favored, similar principles could be employed to design algorithm which "grow" the model starting from a disconnected model and adding edges as appropriate.

More generally, the application of information theory and information geometry to modeling and inference problems is an emerging paradigm in the modern field of graphical modeling. We suspect that there will be much work in this area which will find important applications as graphical modeling methods become more prevalent.

# References

[Aka73] H. Akaike. Information theory and an extension of the maximum likelihood principle. 1973.

[Ama01] S. Amari. Information geometry on hierarchy of probability distributions. *IEEE Transactions on Information Theory*, 47(5):1701–1711, July 2001.

[BN78] O. Barndorff-Nielsen. *Information and Exponential Families*. Wiley series in probability and mathematical statistics. John Wiley, 1978.

[Csi75] I. Csiszár. $I$-divergence geometry of probability distributions and minimization problems. *Annals of Probability*, 3(1):146–158, February 1975.

[Efr78] B. Efron. The geometry of exponential families. *The Annals of Statistics*, 6(2):362–376, 1978.

[Goo63] I.J. Good. Maximum entropy for hypothesis formulation, especially for multidimensional contingency tables. *Annals of Mathematical Statistics*, 34(3):911–934, September 1963.

[Jay57] E.T. Jaynes. Information theory and statistical mechanics, i. *Physics Review*, 106:620–630, 1957.

[Jor99] M.I. Jordan, editor. *Learning in Graphical Models*. Adaptive Computation and Machine Learning Series. The MIT Press, 1999.

[KL51] S. Kullback and R.A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22(1):79–86, March 1951.

[Kul59] S. Kullback. *Information Theory and Statistics*. John Wiley, 1959. Dover reprint, 1997.

[Lau96] S.L. Lauritzen. *Graphical Models*. Oxford Statistical Science Series. Oxford University Press, 1996.

[Sch78] G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6:461–464, 1978.