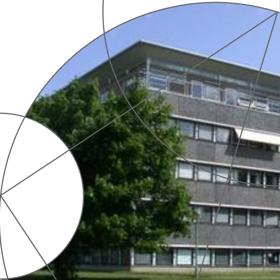
Faculty of Science

Linear estimating equations for Gaussian graphical models with symmetry

Steffen Lauritzen[1]
University of Copenhagen
Department of Mathematical Sciences



useR!, Aalborg, July 2015
Slide 1/31

[1] based on Forbes and Lauritzen (2015).

Motivation

- Højsgaard and Lauritzen (2008) define Gaussian graphical models with symmetry
- maximum likelihood estimation (MLE) is possible but computationally expensive
- search space is huge, so *model selection is difficult,* in particular because of the above
- Is there another way?
- *Yes, use the SME (score matching estimator)!*

Outline

1. Scoring rules
2. Exponential families
3. Gaussian linear concentration models
4. Gaussian graphical models with symmetry
5. Structure identification

Scoring rules

Game between *Forecaster* and *Nature*:

Forecaster quotes probability distribution Q for a random quantity X. Then Nature reveals $X = x$.

How well did Forecaster do? A *score* is calculated $S(x, Q)$ representing a loss to Forecaster.

The function $S(x, Q)$ is a *scoring rule* (Good, 1952; McCarthy, 1956).

A common example of such a scoring rule is the *logarithmic score*

$$S(x, Q) = -\log q(x)$$

where $q(x)$ is the density of Q w.r.t. a fixed measure on $\mathcal{X}$.

We can extend the definition of a scoring rule to $S(P, Q)$ for any probability distribution P as

$$S(P, Q) = \mathbb{E}_{X \sim P}\{S(X, Q)\} = \int S(x, Q)\, P(dx)$$

and further, using the right-hand expression, to $S(\mu, Q)$ for any positive and finite measure. Then S is linear in the first argument.

Proper scoring rules

A scoring rule is *proper* if it encourages honesty, i.e. if the loss is minimized for $Q = P$, i.e. if

$$S(P, P) = \inf_{Q} S(P, Q).$$

It is *strictly proper* if the minimum is unique.

The logarithmic score is strictly proper.

Other examples of strictly proper scoring rules include for $\mathcal{X}$ being finite the *Brier score*

$$S(x, Q) = ||q||_2^2 - 2q(x),$$

where q is the pmf of Q and $||q||_2^2 = \sum_x q(x)^2$, and the *spherical score*

$$S(x, Q) = -q(x)/||q||_2.$$

Also, for $\mathcal{X} = \mathbb{R}$, the *Bregman scores* are strictly proper

$$S(x, Q) = \phi'\{q(x)\} + \int \left[\phi\{q(y)\} - q(y)\phi'\{q(y)\}\right] \mu(dy),$$

where ϕ is any strictly concave real function.

Every strictly proper scoring rule induces an *entropy function*

$$H(P,P) = S(P,P)$$

and a non-negative *divergence* (Dawid, 1998; Grünwald and Dawid, 2004)

$$D(P,Q) = S(P,Q) - S(P,P) = S(P,Q) - H(P) \geq 0.$$

For the *logarithmic score* we get the *Shannon entropy*

$$H(P) = \mathbb{E}_{X\sim P}\{-\log p(X)\}$$

and the *Kullback–Leibler divergence*

$$D(P,Q)) = \mathbb{E}_{X\sim P}\{-\log q(X) + \log p(X)\} = \mathbb{E}_{X\sim P}\{\log p(X)/q(X)\}.$$

Suppose $\mathcal{X} \subseteq \mathbb{R}^p$ and the density $q = dQ/dx$ of Q satisfies:

$$\mathbb{E}_{X \sim P} \|\nabla \log q(X)\|_p^2 < \infty \text{ for all } P, Q \in \mathcal{P};$$

as well as $q(x) \to 0$ and $\|\nabla q(x)\|_p \to 0$ as x approaches the boundary of $\mathcal{X}$.

Then Hyvärinen (2005) showed that the divergence function

$$D_2(P, Q) = \mathbb{E}_{X \sim P} \|\nabla \log q(x) - \nabla \log p(x)\|_p^2$$

where p is the density of P, is induced by the scoring rule

$$S_2(x, Q) = \frac{1}{2} \|\nabla \log q(x)\|_p^2 + \Delta \log q(x).$$

which is *strictly proper* (Dawid and Lauritzen, 2005).

Let $\mathcal{P} = \{Q_\theta, \theta \in \Theta\}$ and $X^1 = x^1, \ldots, X = x^n$ be a sample in $\mathcal{X}$ with empirical distribution $\hat{P}$.

The *score estimator* of θ is determined as the minimizer

$$\check{\theta} = \arg\min_{\theta \in \Theta} \sum_{i=1}^{n} S(x^i, Q_\theta) = \arg\min_{\theta \in \Theta} \mathbb{E}_{X \sim \hat{P}}\{S(X, Q_\theta)\}.$$

Dawid and Lauritzen (2005) show that *this minimization yields an unbiased estimating equation*

$$\sum_{i=1}^{n} S'(x^i, \theta) = 0,$$

where $S'(x, \theta)$ is the vector of derivatives of $S(x, Q_\theta)$ w.r.t. θ.

Solutions to the score equations are *M-estimators* (Huber, 1964, 1967) — generalized means (Brøns et al.) — and are typically consistent, although rarely efficient.

If $S(x, Q) = -\log q(x)$ is the logarithmic score, the equation is the *likelihood equation* and the score estimator is the *maximum likelihood estimator.*

The score matching estimator

The *score matching estimator* (Hyvärinen, 2005) is the estimator corresponding to the scoring rule

$$S_2(x, Q) = \frac{1}{2} \|\nabla \log q(x)\|_p^2 + \Delta \log q(x).$$

Note that $S_2(x, Q)$ can be calculated if we only know q up to an unknown proportionality factor.

Hence, if $q(x \mid \theta) = c(\theta)h(x, \theta)$, *we do not need to have a simple expression for the normalizing constant* $c(\theta)$ as it disappears by differentiation.

Exponential families

Consider an exponential family $\mathcal{P}$ with densities $q(x \mid \theta)$:

$$\log q(x \mid \theta) = \langle \theta, t(x) \rangle_d - a(\theta) + b(x), \quad \theta \in \Theta.$$

Here $t(x) \in L$ is the canonical sufficient statistic, L is a d-dimensional vector space, $\langle \cdot, \cdot \rangle_d$ an inner product on L, and $\Theta \subseteq L$ is the (convex) canonical parameter space. We get

$$\nabla \log q(x \mid \theta) = D(x)\theta + \nabla b(x)$$

where $D(x) = \nabla t(x)$ is determined by $D(x)\eta = \nabla \langle \eta, t(x) \rangle_d$ for all $\eta \in L$ and further

$$\Delta \log q(x \mid \theta) = \langle \theta, \Delta t(x) \rangle_d + \Delta b(x)$$

with $\Delta t(x)$ given by $\langle \eta, \Delta t(x) \rangle_d = \Delta \langle \eta, t(x) \rangle_d$.

The score matching estimator based on $X^1 = x^1, \ldots, X = x^n$ is determined by the linear(!) estimating equation for θ

$$\sum_{i=1}^{n} D(x^i)^* \left\{ D(x^i)\theta + \nabla b(x^i) \right\} + \Delta t(x^i) = 0,$$

where $D(x^i)^*$ is the transpose of $D(x^i)$.

If $\sum_{i=1}^{n} D(x^i)^* D(x^i)$ is invertible, the score estimation equation has the unique solution

$$\check{\theta}_n = -\left\{ \sum_{i=1}^{n} D(x^i)^* D(x^i) \right\}^{-1} \sum_{i=1}^{n} \left\{ D(x^i)^* \nabla b(x^i) + \Delta t(x^i) \right\}.$$

Beware! We may have $\check{\theta}_n \notin \Theta$. Ignore this problem at the moment.

Gaussian linear concentration models

Gaussian models with *linear structure in the concentration matrix* (Anderson, 1970), are special instances.

Let L be a d-dimensional subspace of $\mathcal{S}^p$, the symmetric $p \times p$ matrices with trace inner product $\langle A, B \rangle_d = \mathrm{tr}(AB)$ and associated *Frobenius norm* $\|A\|_d^2 = \mathrm{tr}(A^2)$. Then

$$\begin{aligned}
\log p(x \mid K) &= \{\log\det(K) - p\log(2\pi) - \langle x, Kx\rangle_p\}/2 \\
&= -\left\langle K, xx^\top \right\rangle_d /2 + \{\log\det(K) - p\log(2\pi)\}/2 \\
&= \left\langle K, -\Pi_L(xx^\top)/2 \right\rangle_d + \{\log\det(K) - p\log(2\pi)\}/2
\end{aligned}$$

are exponential families as above with $\mathcal{X} = \mathbb{R}^p$, $\theta = K$, $b(x) = 0$, and $t(x) = -\Pi_L(xx^\top)/2$, where Π_L is the orthogonal projection onto L in $\mathcal{S}^p$.

We may w.l.o.g. assume $I_p \in \Theta$ and then get

$$D(x)K = -Kx, \quad D(x)^* y = -\Pi_L(xy^\top + yx^\top)/2, \quad \Delta t(x) = -I_p$$

where I_p is the $p \times p$ identity matrix.

If we let $W = n^{-1} \sum_{i=1}^n x^i x^{i\top}$, the score matching equation specializes to

$$\Pi_L(K \circ W) = I_p$$

where $A \circ B = (AB^\top + BA^\top)/2$ is the *Jordan product* (Albert, 1946) of the symmetric matrices A and B.

Suppose that L is closed under the Jordan product or, equivalently, $\Theta = L \cap \mathcal{S}_{+}^{p}$ is closed under inversion (Jensen, 1988). Includes all models determined by group invariance (Andersson, 1975).

For such models the MLE and the score matching estimator (SME) coincide. More precisely:

If the subspace L is a Jordan subalgebra, the score matching estimator is equal to the maximum likelihood estimator and

$$\hat{K} = \check{K} = \{\Pi_L(W)\}^{-1},$$

provided $\Pi_L(W)$ is invertible.

Existence issues

Observing $x = (x^1, \dots, x^n)$, the score matching equation has a unique solution iff the quadratic form

$$D_2(K) = \sum_{i=1}^{n} \|Kx^i\|^2$$

is positive definite on L. If $e^1, \dots, e^d$ is an orthogonal basis for L, the matrix for this quadratic form is $M(x) = \{m_{uv}(x)\}$

$$m_{uv}(x) = \sum_{i=1}^{n} \left\langle e^u x^i, e^v x^i \right\rangle_p = n \operatorname{tr}(e^u W e^v)$$

and hence *D_2 is positive definite if and only if* $\det M(x) > 0$.

This determinant is a polynomial in x; hence *either* $\det M(x) = 0$ *for all* x *or* $\det M(x) > 0$ *almost everywhere* (Okamoto, 1973).

Contrast to the MLE, which can exist with probability strictly between zero and one (Buhl, 1993; Uhler, 2012; Gross and Sullivant, 2014).

If the SME exists, then the MLE also exists, i.e. if $K \to \Pi_L(K \circ W)$ has trivial kernel the MLE exists, but not conversely (Forbes and Lauritzen, 2015).

Even when there is a unique solution $\check{K}$, $\check{K}$ *may not be positive semidefinite*.

Say L is *n-estimable* if there is an $x = (x^1, \ldots, x^n) \in \mathbb{R}^{p \times n}$ such that $\det M(x) > 0$.

For $n \geq p$, W is positive definite with probability one and hence $M(x)$ is positive definite and any L is n-estimable.

Assume $n < p$. Let $r = p - n$ and $T_k = k(k+1)/2$.

If $\dim L > T_p - T_r$, *L is not n-estimable.*

The converse is false:

$$L = \left\{ \begin{pmatrix} a & c & 0 & f \\ c & b & -f & 0 \\ 0 & -f & a & c \\ f & 0 & c & b \end{pmatrix} : a, b, c, f \in \mathbb{R} \right\},$$

is not 1-estimable although we have $p = 4$ and $d = 4$ and thus

$$T_p - T_r = T_4 - T_3 = 4 = d.$$

This is an example of a Jordan subalgebra (Jensen, 1988) and — as Jensen — we conclude that also the MLE fails to exist.

Gaussian graphical models with symmetries (Højsgaard and Lauritzen, 2008) are linear concentration models generated by a coloured graph.

Undirected graph $\mathcal{G} = (V, E)$.

Colouring vertices of $\mathcal{G}$ with different colours induces partitioning of V into *vertex colour classes.*

Colouring edges E partitions E into disjoint *edge colour classes*

$$V = V_1 \cup \cdots \cup V_T, \quad E = E_1 \cup \cdots \cup E_S.$$

$\mathcal{V} = \{V_1, \ldots, V_T\}$ is a *vertex colouring,*

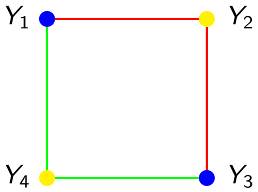
$\mathcal{E} = \{E_1, \ldots, E_S\}$ is an *edge colouring,*

$\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is a *coloured graph.*

RCON model

1. *Diagonal elements K corresponding to vertices in the same vertex colour class must be identical.*
2. *Off–diagonal entries of K corresponding to edges in the same edge colour class must be identical.*

The set of positive definite matrices which satisfy these restrictions is denoted $\mathcal{S}^+(\mathcal{V},\mathcal{E})$.



Corresponding RCON model will have concentration matrix

$$K = \begin{pmatrix} k_{11} & k_{12} & 0 & k_{14} \\ k_{21} & k_{22} & k_{23} & 0 \\ 0 & k_{32} & k_{33} & k_{34} \\ k_{41} & 0 & k_{43} & k_{44} \end{pmatrix}$$

Determines linear concentration model.

Let e^u for $u \in \mathcal{V}$ denote the $|V| \times |V|$ diagonal matrix with $e^u_{\alpha\alpha} = 1$ if $\alpha \in u$ and 0 otherwise. Similarly, for each edge colour class $u \in \mathcal{E}$ we let e^u be the $|V| \times |V|$ symmetric matrix with $e^u_{\alpha\beta} = 1$ if $\{\alpha, \beta\} \in u$ and 0 otherwise. Then $\{e^u, u \in \mathcal{V} \cup \mathcal{E}\}$ form an orthogonal basis for L.

Likelihood equations (Højsgaard and Lauritzen, 2008) become

$$\mathrm{tr}(e^u W) = \mathrm{tr}(e^u K^{-1}), \quad u \in \mathcal{V} \cup \mathcal{E}, \tag{1}$$

which are non-linear in K.

The score matching equations for RCON models are

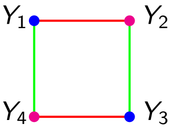
$$\operatorname{tr}(e^u WK) = \operatorname{tr}(e^u), \quad u \in \mathcal{V} \cup \mathcal{E}, \tag{2}$$

which should be compared to (1); they are analogous to the Yule–Walker equations for estimating parameters of autoregressive processes in time series.

Using previous result we find that *the SME does not exist if* $|\mathcal{V}| + |\mathcal{E}| > n(2|V| - n + 1)/2$.

Modify Jordan counterexample to coloured graphical model:

$$L = \left\{ \begin{pmatrix} a & c & 0 & f \\ c & b & f & 0 \\ 0 & f & a & c \\ f & 0 & c & b \end{pmatrix} : a, b, c, f \in \mathbb{R} \right\},$$



This is 1-estimable as $\det M(x) = 4x_1x_2x_3x_4$.

This is not a Jordan subalgebra but *we conclude that also the MLE exists.*

Results of Gross and Sullivant (2014) imply partial results for uncoloured graphs:

The *r-core* of a graph $\mathcal{G}$ is obtained by successively deleting vertices of degree $< r$.

If $\mathcal{G}$ has empty r-core, it is n-estimable for $n \geq r$.

For planar graphs, four observations suffice:

If $\mathcal{G}$ is planar, it is n-estimable for all $n \geq 4$.

Minimum score for the SME is very easy to calculate

$$\sum_{i=1}^{n} S_2(y_i, Q_{\check{K}}) = \operatorname{tr} \check{K}^2 W/2 - n \operatorname{tr}(\check{K}) = -n \operatorname{tr}(\check{K})/2.$$

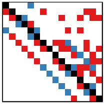
This makes sense even if $\check{K}$ is not positive definite.

So identify graph by minimizing a penalised version, say:

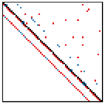
$$\tilde{S}(\mathcal{G}) = (|V| + |E|)\sqrt{p} \log\log(np)/(2n) - \operatorname{tr}(\check{K}_{\mathcal{G}}).$$

This is *extremely fast*. For example, using this on an $s \times s$ lattice so $p = s^2$ it took for $s = 100$, i.e. $p = 10000$ and $n = 100000$ 10 seconds to identify the lattice structure (correctly). Note concentration matrix is 10000×10000, so is rather big...

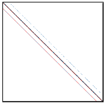
Could not load the concentration matrix into R to compare with, say, graphical lasso.



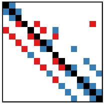
: $p = 16, n = 1 \times p$



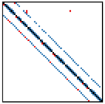
: $p = 64, n = 1 \times p$



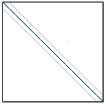
: $p = 256, n = 1 \times p$



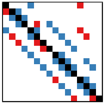
: $p = 16, n = 5 \times p$



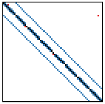
: $p = 64, n = 5 \times p$



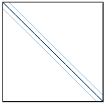
: $p = 256, n = 5 \times p$



: $p = 16, n = 10 \times p$



: $p = 64, n = 10 \times p$



: $p = 256, n = 10 \times p$

Issues to be considered

- Find general condition for existence of the SME ($\det M(x) > 0$);
- When is the SME positive definite?
- When is the SME positive definite with high probability?
- Define fast model screening procedure for structure identification.
- Are there other interesting exponential families where the SME could be used with advantage?
- *Make all this available in R, please...*

Albert, A. A. (1946). On Jordan algebras of linear transformations. *Transactions of the American Mathematical Society*, 59:524–555.

Anderson, T. W. (1970). Estimation of covariance matrices which are linear combinations or whose inverses are linear combinations of given matrices. In Bose, R. C., Chakravarti, I. M., Mahalanobis, P. C., Rao, C. R., and Smith, K. J. C., editors, *Essays in Probability and Statistics*, pages 1–24. University of North Carolina Press, Chapel Hill, N.C.

Andersson, S. A. (1975). Invariant normal models. *The Annals of Statistics*, 3:132–154.

Buhl, S. (1993). On the existence of maximum likelihood estimators for graphical Gaussian models. *Scandinavian Journal of Statistics*, 20:263–270.

Dawid, A. P. (1998). Coherent measures of discrepancy, uncertainty and dependence, with applications to Bayesian predictive experimental design. Technical Report 139, Department of Statistical Science, University College London.

Dawid, A. P. and Lauritzen, S. L. (2005). The geometry of decision theory. In *Proceedings of the Second International Symposium on Information Geometry and its Applications*, pages 22–28. University of Tokyo.

Forbes, P. G. M. and Lauritzen, S. (2015). Linear estimating equations for exponential families with application to Gaussian linear concentration models. *Linear Algebra and its Applications*, 473:261–283.

Good, I. J. (1952). Rational decisions. *Journal of the Royal Statistical Society, Series B*, 14:107–114.

Gross, E. and Sullivant, S. (2014). The maximum likelihood threshold of a graph. arXiv:1404.6989.

Grünwald, P. D. and Dawid, A. P. (2004). Game theory, maximum entropy, minimum discrepancy, and robust

Bayesian decision theory. *Annals of Statistics*, 32:1367–1433.

Højsgaard, S. and Lauritzen, S. L. (2008). Graphical Gaussian models with edge and vertex symmetries. *Journal of the Royal Statistical Society, Series B*, 70:1005–1027.

Huber, P. J. (1964). Robust estimation of a location parameter. *Annals of Applied Statistics*, 35(1):73–101.

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In Cam, L. M. L. and Neyman, J., editors, *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 221–223, Berkeley, CA. University of California Press.

Hyvärinen, A. (2005). Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6:695–709.

Jensen, S. T. (1988). Covariance hypotheses which are linear in both the covariance and the inverse covariance. *The Annals of Statistics*, 116:302–322.

McCarthy, J. (1956). Measures of the value of information. *Proceedings of the National Academy of Sciences*, 42:654–655.

Okamoto, M. (1973). Distinctness of the eigenvalues of a quadratic form in a multivariate sample. *The Annals of Statistics*, 1:763–765.

Uhler, C. (2012). Geometry of maximum likelihood estimation in Gaussian graphical models. *Annals of Statistics*, 40:238–261.