Model selection in penalized Gaussian graphical models

Ernst Wit
University of Groningen

e.c.wit@rug.nl
http://www.math.rug.nl/∼ernst

January 2014

Penalized likelihood generates a PATH of solutions

- Consider an experiment: $|\Gamma|$ genes measured across $|T|$ time points.
- Assume n iid samples $\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(n)}$, where $\mathbf{y}^{(i)} = (y_1^{(i)}, \ldots, y_{\Gamma T}^{(i)})$.
- Assume $\mathbf{Y}^{(i)} \sim N(\mathbf{0}, \boldsymbol{K}^{-1})$, then

Likelihood:

$$l(\boldsymbol{K}|\mathbf{y}) \propto \frac{n}{2}\left\{\log(|K|) - \mathrm{tr}(\mathbf{S}\boldsymbol{K})\right\}.$$

AIM: Optimization of penalized likelihood:

$$\hat{\boldsymbol{K}}_\lambda := \mathrm{argmax}_{\boldsymbol{K}}\{l(\boldsymbol{K}|\mathbf{y})\}$$

subject to

- $\boldsymbol{K} \succeq 0$;
- $||\boldsymbol{K}||_1 \leq 1/\lambda$ ... **for λ in some range!!**;
- some factorial colouring F.

Between proximity and truth

True process for data Y:

$$Y \sim \mathcal{Q}.$$

A *statistical model* is a collection of measures:

$$\mathcal{M}_i = \{\mathcal{P}_\theta \mid \theta \in \Theta_i\}$$

... and typically we consider several: $\mathbb{M} = \{\mathcal{M}_1, \ldots, \mathcal{M}_k\}$.

What is the best model?

- **"Proximity":** the model that is closest to the truth:

$$\min_{i \leq k} KL(\mathcal{P}_{\hat{\theta}_i}; \mathcal{Q}).$$

- **"Truth":** the model that is most likely to be the truth:

$$\max_{i \leq k} P(\mathcal{M}_i | Y).$$

Truth

With flat prior on $\Theta_{\mathcal{M}}$ and $\mathcal{M}$, model probability for $\mathcal{M} \in \mathbb{M}$ is:

$$\begin{aligned}
p(\mathcal{M}|y) &\propto p(y|\mathcal{M})/p(y) \\
&\propto \int_{\Theta_{\mathcal{M}}} e^{n\bar{\ell}(\theta)} \partial\theta \\
&\approx e^{\ell(\hat{\theta})} \int_{\Theta_{\mathcal{M}}} e^{-\frac{1}{2}(\theta-\hat{\theta})^t n \frac{\partial^2}{\partial\theta^2} \bar{\ell}(\hat{\theta})(\theta-\hat{\theta})} \partial\theta \\
&\approx e^{\ell(\hat{\theta})} (2\pi)^{-p/2} n^{-p/2} \left| \frac{\partial^2}{\partial\theta^2} \bar{\ell}(\hat{\theta}) \right|^{-1/2},
\end{aligned}$$

where $p = dim(\Theta_{\mathcal{M}})$ and $\bar{\ell}(\hat{\theta}) = \ell(\hat{\theta})/n$.

Schwarz (1978) ignored terms not depending on n:

$$p(\mathcal{M}|y) \approx e^{l(\hat{\theta})} n^{-p/2}, \quad \text{for large } n.$$

BIC: Bayesian information criterion

By applying the $-\log$ and ignoring constant terms c:

$$BIC(\mathcal{M}) = -2l(\hat{\theta}) + p\log(n).$$

Minimizing BIC corresponds to maximizing posterior model probability.

Define model weights $W(\mathcal{M})$,

$$W(\mathcal{M}) = e^{-BIC(\mathcal{M})/2},$$

which rescaled correspond to posterior model probabilities,

$$p(\mathcal{M}_i|y) = \frac{W(\mathcal{M}_i)}{\sum_{\mathcal{M}\in\mathbb{M}} W(\mathcal{M})}.$$

BIC for Gaussian graphical models

The BIC for an estimated Gaussian graphical model $\widehat{K}$:

$$BIC(\widehat{K}) = n(-\log|\widehat{K}| + Tr(S\widehat{K})) + p_{\widehat{K}} \log(n),$$

where

$$p_K = |\{\text{unique non-zeroes in } K\}|,$$

which is less than $p(p-1)/2$.

BIC for Gaussian graphical models

The BIC for an estimated Gaussian graphical model $\widehat{K}$:

$$BIC(\widehat{K}) = n(-\log|\widehat{K}| + Tr(S\widehat{K})) + p_{\widehat{K}} \log(n),$$

where

$$p_K = |\{\text{unique non-zeroes in } K\}|,$$

which is less than $p(p-1)/2$.

But...

- BIC is an asymptotic approximation. What is n??
- For penalized estimate $\widehat{K}$ are number of parameters not smaller??

eBIC: extended Bayesian information criterion

The BIC for an estimated Gaussian graphical model $\widehat{K}$:

$$eBIC_\gamma(\widehat{K}) = n(-\log|\widehat{K}| + Tr(S\widehat{K})) + p_{\widehat{K}}\log(n) + 4\gamma p_{\widehat{K}}\log(p),$$

where

$$\begin{aligned} p_K &= |\{\text{unique non-zeroes in } K\}| \\ p &= \text{number of nodes} \\ \gamma &= \text{tuning paramter}(0 \leq \gamma \leq 1) \end{aligned}$$

If

$$\begin{aligned} \gamma &= 0 \quad \Rightarrow \text{ordinary BIC} \\ \gamma &= 1 \quad \Rightarrow \text{additional sparsity} \\ \gamma &= 0.5 \Rightarrow \text{good trade-off (Foygel \& Drton, 2010)} \end{aligned}$$

Proximity

Let the true density of the data Y be:

$$Y \sim q = d\mathcal{Q}.$$

The Kullback-Leibler (KL) divergence between fitted and true model

$$KL(\hat{\theta}_i) \;=\; \int q(y) \log q(y) dy - \int q(y) \log p(y; \hat{\theta}_i) dy$$

Proximity

Let the true density of the data Y be:

$$Y \sim q = d\mathcal{Q}.$$

The Kullback-Leibler (KL) divergence between fitted and true model

$$\begin{aligned} KL(\hat{\theta}_i) &= \int q(y) \log q(y) dy - \int q(y) \log p(y; \hat{\theta}_i) dy \\ &= C - E_q(\ell(\hat{\theta}_i) \end{aligned}$$

Proximity

Let the true density of the data Y be:

$$Y \sim q = d\mathcal{Q}.$$

The Kullback-Leibler (KL) divergence between fitted and true model

$$\begin{aligned} KL(\hat{\theta}_i) &= \int q(y) \log q(y) dy - \int q(y) \log p(y; \hat{\theta}_i) dy \\ &= C - E_q(\ell(\hat{\theta}_i)) \\ &\approx C - \ell(\hat{\theta}_i) + p_i^* \end{aligned}$$

where

- $p(.;\theta)$ is density associated with $\mathcal{P}_\theta$.
- $p_i^* = \text{Trace}\left(J_i^{-1} K_i\right) \approx \dim(\Theta_i)$,
  where J_i and K_i are Fisher informations using model $\mathcal{M}_i$:

$$J_i = E_g\left(\frac{\partial^2 \log f(Y, \hat{\theta}_i)}{\partial\theta\partial\theta^t}\right), \quad K_i = V_g\left(\frac{\partial \log(f(Y, \hat{\theta}_i))}{\partial\theta}\right).$$

rijksuniversiteit groningen

AIC: Akaike's information criterion

By applying the $-\log$ and ignoring constant terms c:

$$AIC(\mathcal{M}) = -2l(\hat{\theta}) + 2p.$$

Minimizing AIC corresponds to maximizing posterior model probability.

Define *Akaike weights* $W(\mathcal{M})$,

$$W(\mathcal{M}) = e^{-AIC(\mathcal{M})/2},$$

which rescaled correspond to probability weights that add up to one,

$$p(\mathcal{M}_i) = \frac{W(\mathcal{M}_i)}{\sum_{\mathcal{M}\in\mathbb{M}} W(\mathcal{M})}.$$

AIC for Gaussian graphical models

The AIC for an estimated Gaussian graphical model $\widehat{K}$:

$$AIC(\widehat{K}) = n(-\log|\widehat{K}| + Tr(S\widehat{K})) + 2p_{\widehat{K}},$$

where

$$p_K = |\{\text{unique non-zeroes in } K\}|,$$

which is less than $p(p-1)/2$.

AIC for Gaussian graphical models

The AIC for an estimated Gaussian graphical model $\widehat{K}$:

$$AIC(\widehat{K}) = n(-\log|\widehat{K}| + Tr(S\widehat{K})) + 2p_{\widehat{K}},$$

where

$$p_K = |\{\text{unique non-zeroes in } K\}|,$$

which is less than $p(p-1)/2$.

But...

- AIC is an asymptotic approximation. What is n??
- For penalized estimate $\widehat{K}$ are number of parameters not smaller??

In the next slides, we propose three alternatives.

1. Exact AIC

$$KL(K_0||\widehat{K}) = \frac{1}{2}\{Tr(\widehat{K}\Sigma_0) - \log|\widehat{K}\Sigma_0| - p\}$$

Scaling this by 2 and ignoring a constant

$$2KL(K_0||\widehat{K}) \cong -\{\log|\widehat{K}| - Tr(\widehat{K}S)\} + 2 \cdot \frac{1}{2}\{Tr(\widehat{K}(\Sigma_0 - S))\}.$$

We can write this as

$$2KL(K_0||\widehat{K}) \cong -2\ell(\widehat{K}) + 2 \times \frac{1}{2}\{Tr(\widehat{K}(\Sigma_0 - S))\}.$$

Definition (Degrees of freedom in Gaussian graphical model)

Let $Y \sim N(0, K_0^{-1})$ and $\widehat{K}$ an estimate of K_0:

$$\mathrm{df}_{\widehat{K}} = \frac{1}{2}\{Tr(\widehat{K}(\Sigma_0 - S))\}.$$

Approximate Exact AIC

We obviously don't know Σ_0, but we can estimate it:

$$\widehat{\Sigma}_0 = \pi S - (1-\pi)\mathrm{diag}\{\sigma_{11}^2, \ldots, \sigma_{pp}^2\},$$

for some tuning paramter π.

Definition (Approximate Exact AIC)

Let $Y \sim N(0, K_0^{-1})$ and $\widehat{K}$ an estimate of K_0:

$$AIC(\widehat{K}) = -2\ell(\widehat{K}) + 2\widehat{\mathrm{df}},$$

where

$$\widehat{\mathrm{df}} = \frac{1}{2}\{Tr(\widehat{K}(\widehat{\Sigma}_0 - S))\}.$$

2. Exact GIC

Problem of AIC: $\hat{K}_\lambda$ is *not* a MLE.

GIC for M-estimator $\hat{K}$ derived by Konishi & Kitagawa (1996):

$$\mathrm{GIC} = -2\sum_{k=1}^{n} l_k(\hat{K}; x_k) + 2\mathrm{tr}\{R^{-1}Q\}, \tag{1}$$

where R and Q are square matrices of order p^2 given by

$$R = -\frac{1}{n}\sum_{k=1}^{n}\{D\psi(x_k, K)\}^\top\big|_{K=\hat{K}},$$

$$Q = \frac{1}{n}\sum_{k=1}^{n}\psi(x_k, K)Dl_k(K)\big|_{K=\hat{K}}.$$

Problem: Bias term in (1) requires inversion of $d^2 \times d^2$ matrix R.

Approximate GIC (Abbruzzo, Vujacic, Wit)

We derived an explicit estimator of KL that avoids matrix inversion:

$$\widehat{\text{GIC}}(\lambda) = -2l(\hat{K}_\lambda) + 2\widehat{\text{df}}_{\text{GIC}},$$

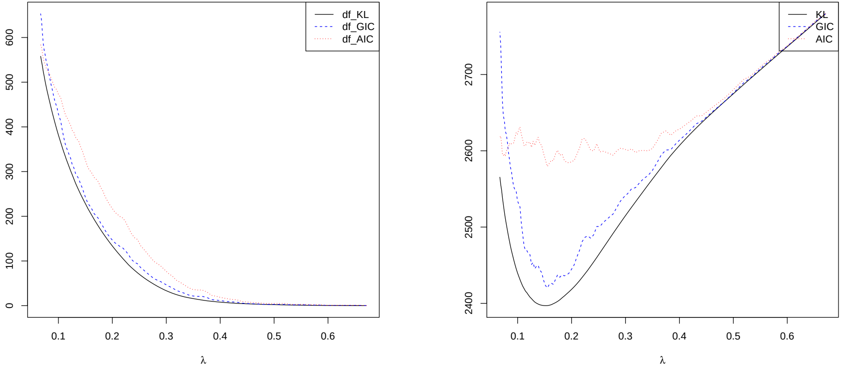
where

$$\widehat{\text{df}}_{\text{GIC}} = \frac{\sum_{k=1}^{n} \text{c}(S_k \circ I_\lambda)^\top \text{c}(\hat{K}_\lambda(S_k \circ I_\lambda)\hat{K}_\lambda)}{2n} - \frac{\text{c}(S \circ I_\lambda)^\top \text{c}(\hat{K}_\lambda(S \circ I_\lambda)\hat{K}_\lambda)}{2},$$

where

- $c()$ is the vectorize operator,
- $S_k = x_k x_k^T$,
- $I_\lambda = 1$ `*` `(`$\hat{K}_\lambda$ `!= 0)`, an indicator matrix.

Approximate GIC: Simulations



Bias term (left) and **KL** divergence (right) estimates for $n = 100, d = 50$.

3. Exact Cross-validation (CV)

Ignoring an additive constant, the KL divergence is:

$$KL(K_0||\widehat{K}) = -\frac{1}{2}\log|\widehat{K}| + E_{\Sigma_0}\frac{1}{2}\{Tr(\widehat{K}XX^t)\}$$

The idea of cross-validation is to replace the final expectation by

$$KL(K_0||\widehat{K}) \approx -\frac{1}{2n}\sum_{k=1}^{n}\{\log|\widehat{K}^{-(k)}| + Tr(\widehat{K}^{-(k)}x_k x_k^t)\}$$

Clearly, this would require recalculating the estimate $\widehat{K}^{-(k)}$ many times.

Approximate Cross-validation (Vujacic, Abbruzzo, Wit)

We write the KL divergence as follows,

$$KL(K_0|\hat{K}_\lambda) = -\frac{1}{n} l(\hat{K}_\lambda) + \text{bias},$$

where $l(K) = n\{\log|K| - \text{tr}(KS)\}/2$ and bias $= \text{tr}(\hat{K}_\lambda(K_0^{-1} - S))/2$.

Definition

LOOCV-inspired estimate of Kullback-Leibler divergence

$$KLCV(\lambda) = -\frac{1}{n} l(\hat{K}_\lambda) + \frac{\sum_{i=1}^n \text{c}[(\hat{K}_\lambda^{-1} - S_i) \circ I_\lambda]^\top (\hat{K}_\lambda \otimes \hat{K}_\lambda) \text{c}[(S - S_i) \circ I_\lambda]}{n(n-1)},$$

where

- $c()$ is the vectorize operator,
- $S_k = x_k x_k^T$,
- $I_\lambda = 1 \texttt{ * } (\hat{K}_\lambda \texttt{ != 0})$, an indicator matrix.

Proof

$$
\begin{aligned}
LOOCV &= -\frac{1}{2n}\sum_{i=1}^{n} f(S_i, \hat{K}^{(-i)}) \\
&= -\frac{1}{2} f(S, \hat{K}) - \frac{1}{2n}\sum_{i=1}^{n}[f(S_i, \hat{K}^{(-i)}) - f(S_i, \hat{K})] \\
&\approx -\frac{1}{n} l(\hat{K}) - \frac{1}{2n}\sum_{i=1}^{n}\left[\frac{df(S_i, \hat{K})}{d\Omega}\right]^{\top} \mathrm{c}(\hat{K}^{(-i)} - \hat{K}).
\end{aligned}
$$

Matrix differential calculus: $df(S_i, \hat{K})/d\Omega = \mathrm{c}(\hat{K}^{-1} - S_i)$.
The term $\mathrm{c}(\hat{K}^{(-i)} - \hat{K})$ is obtained via the Taylor expansion

$$
0 \approx \frac{df(S, \hat{K})}{d\Omega} + \frac{d^2 f(S, \hat{K})}{d\Omega^2}\mathrm{c}(\hat{K}^{(-i)} - \hat{K}) + \frac{d^2 f(S, \hat{K})}{d\Omega dS}\mathrm{c}(S^{(-i)} - S).
$$

Inserting: $df(S, \hat{K})/d\Omega = \mathrm{c}(\hat{K}^{-1} - S)$, $\quad d^2 f(S, \hat{K})/d\Omega dS = -I_{p^2}$,
$d^2 f(S, \hat{K})/d\Omega^2 = -\hat{K}^{-1} \otimes \hat{K}^{-1}$

and consequently

$$
\mathrm{c}(\hat{K}^{(-i)} - \hat{K}) = -(\hat{K} \otimes \hat{K})\mathrm{c}(S^{(-i)} - S).
$$

KLCV simulation results

d=100	KL ORACLE	KLCV	AIC	GACV
n=20	8.06	**8.60**	12.24	28.59
	(0.37)	(0.45)	(0.28)	(19.94)
n=30	6.87	**7.29**	10.59	32.07
	(0.34)	(0.39)	(0.41)	(2.77)
n=50	5.24	**5.63**	7.33	16.93
	(0.27)	(0.33)	(0.81)	(1.40)
n=100	3.34	**3.57**	3.63	6.81
	(0.19)	(0.23)	(0.48)	(0.52)
n=400	1.13	1.20	**1.17**	1.24
	(0.07)	(0.08)	(0.08)	(0.07)

Conclusions

- BIC aims to find true model
- AIC aims to come closest to the truth
- BIC gives sparser model than AIC (typically)
- AIC/BIC have asymptotic issues.
- What are number of parameters for penalized inference?
- Improved versions are available!