Modified Information Criteria in High-Dimensional Data

Xin Gao, York University, Toronto
Joint work with Peter Song and Amy Wu
June, 2011

Outline of the talk

- Information Criterion based on pseudo-likelihood with application to select mean structure
- Information Criterion based on penalized-likelihood with application to select covariance structure

Composite likelihood methodology

- The composite likelihood (CL) paradigm (Lindsay, 1988; Cox and Reid, 2004) constitutes a rich class of pseudo-likelihoods based on marginal likelihood objects.
- Let $\{f(\mathbf{y};\psi), \psi \in \Psi\}$ be a parametric statistical model, with the parameter space $\Psi \subseteq \mathcal{R}^Q$. Let $\mathbf{Y} = (\mathbf{Y}_1', \ldots, \mathbf{Y}_n')'$ denote the data set, where $\mathbf{Y}_i = (y_{i1}, \ldots, y_{im_i})'$ are the vector of observations sampled independently on unit i, $i = 1, \ldots, n$.

Composite likelihood methodology

- $\psi = (\theta, \eta)$, where θ is the parameter of interest and η is the nuisance parameter.
- Model selection in CL is concerned with θ, and the corresponding parameter space is $\Theta \subseteq \mathcal{R}^P$, with dimension P possibly dependent on the sample size.
- A collection of index subsets $\mathcal{A} = \{A : A \subseteq \Omega\}$, where each element A is a subset of $\Omega = \{(i,j), j = 1, \ldots, m_i, i = 1, \ldots, n\}$. For a given unit i, similarly we denote $\mathcal{A}_i = \{A : A \subseteq \Omega_i\}$ with $\Omega_i = \{(i,j), j = 1, \ldots, m_i\}$.
- $\mathbf{Y}_A = \{y_{ij}, (i,j) \in A\}$.

Composite likelihood methodology

- A composite likelihood function is defined as

$$\mathrm{CL}(\boldsymbol{\theta}; \mathbf{Y}) = \prod_{A \in \mathcal{A}} L_A(\boldsymbol{\theta}; \mathbf{Y})^{w_A} = \prod_{i=1}^{n} \prod_{A \in \mathcal{A}_i} L_A(\boldsymbol{\theta}; \mathbf{Y})^{w_A}, \quad (1)$$

where $L_A(\boldsymbol{\theta}; \mathbf{Y}) = f(\mathbf{Y}_A; \boldsymbol{\theta})$ is the marginal likelihood with respect to composite set A, and $\{w_A\}$ is a set of suitable weights.

Composite likelihood methodology

- Example 1: A singleton $\mathcal{A}_i = \{\Omega_i\}$ corresponds to the full likelihood
- Example 2: $\mathcal{A}_i = \{\{1\}, \ldots, \{m_i\}\}$ gives rise to a composite likelihood of univariate margins.
- The composite log-likelihood is $\mathrm{cl}(\boldsymbol{\theta}; \mathbf{Y}) = \sum_{i=1}^{n} \sum_{A \in \mathcal{A}_i} w_A \ell_A(\boldsymbol{\theta}; \mathbf{Y})$, where $\mathrm{cl}(\boldsymbol{\theta}; \mathbf{Y}) = \log \mathrm{CL}(\boldsymbol{\theta}; \mathbf{Y})$ and $\ell_A(\boldsymbol{\theta}; \mathbf{Y}) = \log L_A(\boldsymbol{\theta}; \mathbf{Y})$.

Composite likelihood methodology

- The maximum composite likelihood estimator (CLE) is given by

$$\hat{\theta}^{c} = \arg\max_{\theta \in \Theta} \mathrm{cl}(\theta; \mathbf{Y}).$$

- The CLE is consistent and asymptotically normally distributed under some mild regularity conditions.

Bayesian Information Criteria

- Let $P = \dim(\Theta)$, and let s be a subset of $\{1, \ldots, P\}$. Denote by θ_s the parameter θ with those elements outside s being pre-specified as 0 or some known values.
- Let d_s be the number of parameters under a marginal submodel s. Let $\mathcal{S}$ denote the model space of all possible submodels being considered. Associated with each submodel s, let $p(s)$ be the prior probability of the occurrence of the submodel defined on space $\mathcal{S}$.

Bayesian Information Criteria

- In the conventional setting where the number of parameters P is fixed (or not dependent on the sample size n), it is commonly assumed that each submodel s has an equal probability of being selected, $p(s) = 1/\text{card}(\mathcal{S})$.
- Under the full likelihood framework, assuming equal priors for different submodels, Schwarz (1978) proposed the BIC criterion to select the best model among all the candidate models. The first term in BIC is minus twice the log-likelihood evaluated at the maximum likelihood estimate and the second term is $\log(n)$ times the number the parameters in the model.

Bayesian Information Criteria

- A much more challenging task of model selection in high-dimensional data analysis is that P is not fixed but increases as the sample size rises. Suppose that $P = O(n^{\kappa})$, with $\kappa > 0$.
- In this case, the equal probability prior will actually favor models with more parameters; see for example Chen and Chen (2008).
- To ensure an increasing chance of selecting models with sparsity, we adopt a stratified sampling scheme proposed by Chen and Chen (2008).

Bayesian Information Criteria

- To proceed, first partition the model space into submodel spaces $\mathcal{S} = \cup_{k=1}^{P} \mathcal{S}_k$, where each $\mathcal{S}_k$ contains models with k parameters.
- Let $\tau(\mathcal{S}_k) = \text{card}(\mathcal{S}_k)$ be the size of $\mathcal{S}_k$. Obviously, $\tau(\mathcal{S}_1) = P$.
- Within a given subspace $\mathcal{S}_k$, an equal probability prior is imposed as $p(s|\mathcal{S}_k) = 1/\tau(\mathcal{S}_k), s \in \mathcal{S}_k$.
- Specifying prior probabilities for these subspaces proportional to their sizes, say $p(\mathcal{S}_k) \propto \{\tau(\mathcal{S}_k)\}^{\xi}$ for some $\xi \leq 1$, we obtain that the prior probability of a submodel s is proportional to $\tau(\mathcal{S}_k)^{-\gamma}$, with $\gamma = 1 - \xi > 0$.
- Using such prior probabilities, Chen and Chen (2008) have proposed an extended BIC criterion which has an extra penalty term $2\gamma \log \tau(\mathcal{S}_k)$ for $s \in \mathcal{S}_k$ on the model space complexity.

Bayesian Information Criteria

- When the full likelihood is numerically prohibitive to compute, we aim to develop an analogue of extended BIC criterion based on the composite likelihood.
- Select the model with the highest composite posterior probability.

$$P_c(s|\mathbf{Y}) = \frac{p(s)\int \mathrm{CL}(\mathbf{Y}|\theta_s)\pi_s(\theta_s)d\theta_s}{\sum_{s\in\mathcal{S}} p(s)\int \mathrm{CL}(\mathbf{Y}|\theta_s)\pi_s(\theta_s)d\theta_s},$$

with $\pi_s(\theta_s)$ denoting the prior density of θ_s.

- Using the Laplace approximation (Tierney and Kadane, 1986, Tierney, et al., 1989), and ignoring $O_p(1)$ terms, we have the resulting criterion simplified as:

$$-2\log \mathrm{CL}(\hat{\theta}_s^c; \mathbf{Y}) + d_s \log(n) + 2\gamma \log\{\tau(\mathcal{S}_{d_s})\}. \quad (2)$$

Measure of model complexity

- $d_s^* = \text{trace}\,(\mathbf{H}_s^{-1}\mathbf{V}_s)$, where

$$\mathbf{H}_s = \mathrm{E}_{\psi_{T,0}}\left\{-\tilde{\mathbf{cl}}^{(2)}(\theta_s)\right\}, \text{ and } \mathbf{V}_s = \mathrm{var}_{\psi_{T,0}}\left\{\tilde{\mathbf{cl}}^{(1)}(\theta_s)\right\}. \tag{3}$$

- The d_s^* has been accepted as a measure of model complexity in composite likelihood setting (Varin and Vidoni, 2005).
- The proposed CL-BIC for model selection is:

$$\text{CL-BIC}(s) = -2\log \text{CL}(\hat{\theta}_s^c; \mathbf{Y}) + d_s^* \log(n) + 2\gamma \log\{\tau(\mathcal{S}_{d_s^*})\}, \tag{4}$$

with the cardinality term $\tau(\mathcal{S}_{d_s^*}) = P^{d_s^*}$.

Model Selection Consistency

- The notion of consistent model selection is about identifying the smallest correct model with probability tending to one as the sample size increases.
- Let CL-BIC(s), $s = T, s-, s+$ denote the composite likelihood BIC criteria obtained under the true (T), under-fitting ($s-$) and over-fitting marginal models ($s+$).
- We assume the conventional regularity conditions required for consistency and asymptotic normality of the maximum likelihood estimator (Cox and Hinkley, 1974). Furthermore, we assume several additional regularity conditions needed by composite likelihood estimation in connection to model misspecification (White, 1982; Varin and Vidoni, 2005).

Assumptions

- Denote the composite log-likelihood ratio (CLR) between two marginal submodels s and s' by

$$\lambda_{s'|s}(\mathbf{Y};\boldsymbol{\theta}_{s'},\boldsymbol{\theta}_s) = \log\left\{\frac{\mathrm{CL}(\boldsymbol{\theta}_{s'};\mathbf{Y})}{\mathrm{CL}(\boldsymbol{\theta}_s;\mathbf{Y})}\right\} = \mathrm{cl}(\boldsymbol{\theta}_{s'};\mathbf{Y}) - \mathrm{cl}(\boldsymbol{\theta}_s;\mathbf{Y}). \tag{5}$$

- The pseudo true value of parameter $\boldsymbol{\theta}_s$ in Θ_s under a misspecified model s, which minimizes the expected composite KL distance (Varin and Vidoni, 2005) between the true marginal model and a marginal submodel s. That is,
$\boldsymbol{\theta}_{s,0} = \arg\min_{\boldsymbol{\theta}_s\in\Theta_s} \mathrm{E}_{\boldsymbol{\psi}_{T,0}}\{\lambda_{T|s}(\mathbf{Y};\boldsymbol{\theta}_{T,0},\boldsymbol{\theta}_s)\}$.

Model identifiability

- To establish the consistency result, we need a set of regularity assumptions regarding the uniform boundedness of the moments of the derivatives of the composite log-likelihood across the model space.
- Between the true model and a competing model s, we examine the standardized expected composite KL distance:

$$\mathrm{E}_{\psi_{T,0}}\{\lambda_{T|s}(\mathbf{Y};\boldsymbol{\theta}_{T,0},\boldsymbol{\theta}_{s,0})\}/[\mathrm{var}_{\psi_{T,0}}\{\lambda_{T|s}(\mathbf{Y};\boldsymbol{\theta}_{T,0},\boldsymbol{\theta}_{s,0})\}]^{\frac{1}{2}}.$$

- To ensure model identifiability, we assume that as n increases, the minimum standardized expected KL distance should increase at a rate greater than $\sqrt{\log n}$.

Model identifiability

- Example: in the linear model setting. Consider a model $\mathbf{Y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim N_n(0, \sigma^2\mathbf{I})$. Let $\mathbf{X}_T$ and $\mathbf{X}_s$ denote the design matrices of the true model and a candidate model with respective vectors of the regression coefficients $\boldsymbol{\theta}_T$ and $\boldsymbol{\theta}_s$. Denote the true null value as $\boldsymbol{\theta}_{T,0}$ under the true model and the pseudo null value as $\boldsymbol{\theta}_{s,0}$ under the candidate model. Then Assumption reduces to the condition in Chen and Chen (2008):

$$\lim_{n\to\infty} \min_{s\in S_-} \left\{ (\log n)^{-1} \Delta_n(s) \right\} = \infty, \tag{6}$$

with $\Delta_n(s) = ||\mathbf{X}_T\boldsymbol{\theta}_T - \mathbf{X}_s(\mathbf{X}_s'\mathbf{X}_s)^{-1}\mathbf{X}_s'\mathbf{X}_T\boldsymbol{\theta}_{T,0}||_2^2$.

Some notation

- For over-fitting scenario, define the model space $S_+(m) \subset S_+$, with $S_+(m) = \{s : s \in S_+, d_s - d_T = m\}$, $m = 1, \ldots, K - d_T$.
- For any over-fitting model s, define a matrix $\mathbf{D}_s = (\mathbf{I}_{d_T}, \mathbf{0}_{d_T, d_s - d_T})$, with $\mathbf{I}_{d_T}$ being an identity matrix of dimension $d_T \times d_T$, and $\mathbf{0}_{d_T, d_s - d_T}$ denoting a matrix of zeros with dimension $d_T \times (d_s - d_T)$.
- Let $\mathbf{M}_{s/T}$ denote the difference matrix $(\mathbf{H}_s(\theta_{s,0})^{-1} - \mathbf{D}_s' \mathbf{H}_T^{-1}(\theta_{T,0}) \mathbf{D}_s)$.

Some notation

- let $\lambda_{s[1]}, \ldots, \lambda_{s[m]}$ denote the nonzero eigenvalues of $\mathbf{M}_{s/T}^{\frac{1}{2}} \mathbf{V}_s(\theta_{s,0}) \mathbf{M}_{s/T}^{\frac{1}{2}}$ in ascending order and $\overline{\lambda}_s = \sum_{j=1}^{m} \lambda_{s[j]}/m$. Define $\varpi = \limsup_{n\to\infty} \max_{s\in S_+} (\lambda_{s[m]}/\overline{\lambda}_s)$.
- When all the eigenvalues are equal, the ratio of the maximum eigenvalue over the mean eigenvalue, $\lambda_{s[m]}/\overline{\lambda}_s$, is one. On the other hand, $\lambda_{s[m]}/\overline{\lambda}_s < m$. Thus ϖ resides in interval $[1, K - d_T)$.

Selection consistency

- **Theorem** Under the regularity conditions, when $\gamma > \varpi - 1/(2\kappa)$,

$$P_{\psi_{T,0}}\Big\{ \min_{s \in S_- \cap S_+} \text{CL-BIC}(s) > \text{CL-BIC}(T) \Big\} \to 1,$$

as $n \to \infty$.

Multivariate normal model

- We consider the multivariate familial data analysis discussed in Zhao and Joe (2005). The sample is drawn from families with inter-correlations among individuals in a family. Denote the numbers of families and members in each family by n and m.
- The response vector of measurements for the i-th family is denoted by $\mathbf{Y}_i = (y_{i1}, \ldots, y_{im})'$. Associated is a set of covariates at the individual level, $\mathbf{X}_i = (\mathbf{x}_{i1}, \ldots, \mathbf{x}_{im})'$, with $\mathbf{x}_{ik} = (\mathbf{x}_{ik1}, \ldots, \mathbf{x}_{ikP})'$, representing the P covariates observed for the k-th individual in the i-th family.

Multivariate normal model

- $\mathbf{Y}_i$ follows a multivariate normal distribution, $N_m(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$, where the mean vector is governed by a linear model, $\boldsymbol{\mu}_i = \mathbf{X}_i\boldsymbol{\beta}$, with $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_P)'$. The covariance matrix $\boldsymbol{\Sigma}$ is specified according to an exchangeable dependence structure, $\sigma_{k,k'} = \rho$.
- We consider two different scenarios. In the first scenario, we set $P = 30$, $n = 200$ and $m = 4$. The covariates are generated from a multivariate normal with the standard normal $N(0, 1)$ marginals and inter-correlation $\text{Cov}(x_{ikp}, x_{ikp'}) = 0.2$. The within-family correlation ρ is set to either 0.3 or 0.6.
- In the second scenario, we set $P = 1000$, $n = 200$, and $m = 4$.

Multivariate normal model

Table: Multivariate normal model with $P = 30$ and $N = 200$

β	ρ_y	CL-AIC	CL_U-BIC_0	CL_U-$BIC_{0.5}$	CL_B-BIC_0	CL_B-$BIC_{0.5}$	$EBIC_0$
β_1	0.3	0.878	0.739	0.655	0.911	0.875	0.914
		(0.035)	(0.003)	(0.002)	(0.037)	(0.011)	(0.034)
β_1	0.6	0.873	0.727	0.668	0.946	0.903	0.949
		(0.026)	(0.002)	(0.002)	(0.053)	(0.014)	(0.055)
β_2	0.3	0.852	0.697	0.667	0.892	0.818	0.892
		(0.108)	(0.005)	(0.004)	(0.116)	(0.045)	(0.128)
β_2	0.6	0.845	0.695	0.663	0.938	0.890	0.940
		(0.095)	(0.014)	(0.006)	(0.142)	(0.065)	(0.135)

Multivariate normal model

Table: multivariate normal model with $P = 1000$ and $n = 200$

β	ρ_y	CL-AIC	CL_B-BIC_0	CL_B-$BIC_{0.5}$	$EBIC_0$	$EBIC_{0.5}$
β_1	0.3	0.896	0.893	0.819	0.889	0.818
		0.472	0.439	0.039	0.378	0.037
β_1	0.6	0.894	0.894	0.837	0.881	0.838
		0.456	0.346	0.052	0.211	0.052
β_2	0.3	0.868	0.850	0.717	0.842	0.710
		0.780	0.641	0.044	0.545	0.032
β_2	0.6	0.873	0.847	0.728	0.815	0.722
		0.783	0.535	0.064	0.316	0.053

Multivariate probit model

- The second simulation study is based on a multivariate probit model, in which the binary response vector arises from a dichotomization of an underlying multivariate normally distributed random vector.
- Under the same setup in Section 4.1, binary correlated responses are obtained by dichotomizing the continuous multivariate normal measurements.
- The two scenarios of $P < n$ and $P >> n$ are considered. For a multivariate probit model with many covariates, the full likelihood involves high dimensional integration and is computationally prohibitive.

Multivariate probit model

Table: Multivariate probit model with $P = 30$ and $n = 100$

β	ρ_y	rate	CL-AIC	CL_U-BIC_0	CL_U-$BIC_{0.5}$	CL_B-BIC_0	CL_B-$BIC_{0.5}$
β_1	0.3	(PSR)	0.846	0.710	0.670	0.768	0.682
		(FDR)	0.248	0.068	0.060	0.111	0.063
	0.6	(PSR)	0.850	0.713	0.675	0.769	0.707
		(FDR)	0.233	0.067	0.052	0.104	0.063
β_2	0.3	(PSR)	0.812	0.693	0.687	0.707	0.692
		(FDR)	0.394	0.079	0.071	0.111	0.078
	0.6	(PSR)	0.813	0.703	0.695	0.735	0.693
		(FDR)	0.363	0.089	0.065	0.130	0.069

Multivariate probit model

Table: Multivariate probit model with $P = 1000$ and $n = 100$

β	ρ_y	CL-AIC	CL_U-BIC_0	CL_U-$BIC_{0.5}$	CL_U-$BIC_{1.0}$	CL_B-BIC_0	CL_B-$BIC_{0.5}$	CL_B-$BIC_{1.0}$
β_1	0.3	0.790	0.766	0.593	0.393	0.782	0.647	0.475
		0.522	0.431	0.118	0.029	0.494	0.169	0.055
	0.6	0.778	0.756	0.571	0.398	0.775	0.640	0.500
		0.540	0.448	0.095	0.024	0.516	0.181	0.058
β_2	0.3	0.865	0.840	0.635	0.540	0.863	0.692	0.588
		0.703	0.587	0.092	0.012	0.696	0.163	0.031
	0.3	0.868	0.828	0.637	0.518	0.858	0.718	0.592
		0.711	0.590	0.087	0.012	0.678	0.167	0.036

Quadratic Exponential Model

- A less simple model involving conditional likelihoods.
- Consider an experiment involving n clusters, the i-th of which contains n_i binary measurements. Suppose $y_{ij} = 1$ when the outcome is success and $y_{ij} = -1$ when the outcome is failure.
- Let $\mathbf{Y}_i$ represent the vector of outcomes for the i-th cluster. Geys, et al. (1997) used the following model for the joint distribution of clustered binary data:

$$f_{Y_i}(y_i) \propto \exp\{\sum_{j=1}^{n_i} \mu_{ij} y_{ij} + \sum_{j \leq j'} w_{ijj'} y_{ij} y_{ij'}\}, \qquad (7)$$

which belongs to the quadratic exponential family discussed in Zhao and Prentice (1990).

Quadratic Exponential Model

- Express the joint distribution in terms of z_i, the number of successes for the ith cluster. Assuming $\mu_{ij} \equiv \mu_i$ and $w_i \equiv w$, and through re-parametrization $\mu_i^* = 2\mu_i$, and $w^* = 2w$, model (7) is transformed into:

$$f_{\mathbf{Y}_i}(\mathbf{y}_i) = \exp\{\mu_i^* z_i + w^*(-z_i(n_i - z_i)) - C(\mu_i^*, w^*)\},$$

with $C(\mu_i^*, w^*)$ being the normalizing constant. A positive interaction effect w^* corresponds to classical clustering or over-dispersion, while a negative value corresponds to under-dispersion.

Quadratic Exponential Model

- Using traditional likelihood approach to analyze such data will inevitably involve highly intensive calculation of the normalizing constant $C(\mu_i^*, w^*)$, which varies across clusters of different sizes.
- As an appealing alternative method, we formulate the composite likelihood in the form of conditional likelihoods.

$$\text{cl} = \sum_{i=1}^{n} \sum_{j=1}^{n_i} \log f(y_{ij} | \{y_{ij'}\}, j' \neq j).$$

- Within each cluster, there are n_i conditional probabilities of observing the outcome for the j-th measurement, given the outcome for the other $n_i - 1$ measurements.

Quadratic Exponential Model

- Under the assumption of the exchangeable nature of the measurement, there are two types of contributions: i) the conditional probability of an additional success result, given there are $z_i - 1$ successes and $n_i - z_i$ failures:

$$p_{is} = \frac{\exp\{\mu_i^* - w^*(n_i - z_i + 1)\}}{1 + \exp\{\mu_i^* - w^*(n_i - z_i + 1)\}},$$

ii) the conditional probability of an additional failure, given there are z_i successes and $n_i - z_i - 1$ failures:

$$p_{if} = \frac{\exp\{-\mu_i^* + w^*(n_i - z_i - 1)\}}{1 + \exp\{-\mu_i^* + w^*(n_i - z_i - 1)\}}.$$

Thus, the composite likelihood can be expressed as $\mathrm{cl} = \sum_{i=1}^{n}\{z_i \log p_{is} + (n - z_i)\log p_{if}\}$.

Quadratic Exponential Model

- Modelling in terms of covariate effect can be achieved using the linear model $\mu_i^* = \mathbf{X}_i\beta$, where $\mathbf{X}_i$ is a $1 \times P$ vector containing the covariate values and β is a $P \times 1$ vector of regression coefficients.

Quadratic Exponential Model

Table: Quadratic exponential model with $P = 1000$

n	w	rate	CL-AIC	CL-BIC
500	0.2	(PSR)	0.937	0.933
		(FDR)	0.777	0.218
	0.3	(PSR)	0.921	0.915
		(FDR)	0.742	0.250
	0.4	(PSR)	0.923	0.914
		(FDR)	0.728	0.394
1000	0.2	(PSR)	0.936	0.936
		(FDR)	0.809	0.016
	0.3	(PSR)	0.923	0.923
		(FDR)	0.783	0.032
	0.4	(PSR)	0.914	0.914
		(FDR)	0.757	0.139

Real Data Analysis

- A data from a diabetic nephropathy (DN) study at University of Michigan.
- In this data set, 35 DN abnormal patients were followed for a period of time ranging from 6.91 to 10.89 years. During the period, their renal functions were measured at multiple time points and the treatment results were classified into binary outcomes as either successful or failure.
- Each of the patients' renal tissue had undergone a microarray analysis to obtain gene expression data.

Real Data Analysis

- The purpose: determine if there are any biomarkers among the 500 candidate genes that have important influence on the risk of exacerbation through a certain therapeutic program.
- The challenge:the presence of correlation among repeated measurements within each patient; the number of repeated measurements varies across the 35 patients.
- The total number of measurements is 402, while the total number of candidate covariates is 500.
- This data set is a practical example containing a large number of covariates and strong dependency among the clustered binary outcomes.

Real Data Analysis

- We impose penalization on the composite likelihood with L_1 penalty.
- We gradually increase the tuning parameter in the penalty term and obtain a sequence of models, at which both CL-AIC and CL-BIC are computed.
- We choose $\gamma = 1 - 1/(2\kappa) = 0.75$, setting $\kappa = 2$, and $\hat{\varpi} = 1$, the lower bound of ϖ.
- When the tuning parameter increases from 0.04 to 0.26, the number of parameters in the sequence of selected models are $3, 4, 5, 8, 9, 11$, respectively.

Real Data Analysis

- For the sequence of selected models, the CL-BIC takes the values of $557.85, 551.21, 559.21, 578.96, 585.45, 567.32$, whereas the CL-AIC takes the value of $536.09, 518.5732, 515.7058, 502.8211, 498.4278, 458.5472$.
- As shown, CL-BIC is minimized at an intermediate model with 4 parameters including intercept, interaction and 2 gene covariates. Among all the models being examined, CL-AIC is minimized at the most complicated model with 11 parameters.
- The CL-BIC shows its advantage of balancing the model fitting and model complexity when dealing with a large model space.

Some practical concerns

- One can vary the magnitude of γ and selects the optimum value that offers the best balance between sensitivity and selectivity. This works for the situation when we can simulate the data generating mechanism where the the real data arises and apply the simulation gauged γ onto the real data.
- Another approach uses $\gamma = \hat{\varpi} - 1/(2\kappa)$, under the circumstances that the sample size is sufficient. For each candidate model *s*, we compute the ratio of the maximum eigenvalue over the mean eigenvalue of the matrix $\hat{\mathbf{H}}_s^{-1/2}\hat{\mathbf{V}}_s\hat{\mathbf{H}}_s^{-1/2}$. The maximum ratio over all the models being examined offers an ad-hoc estimator $\hat{\varpi}$ of the quantity ϖ.
- When sample size is small, as a conservative approach, we choose $\gamma = 1 - 1/(2\kappa)$ setting $\hat{\varpi} = 1$, the lower bound of ϖ.

model selection in covariance structure

- Covariance selection in Gaussian graphical model
- The proposed penalized likelihood method for covariance selection
- Selection of Tuning parameter –Consistency of modified BIC in penalized likelihood framework

Covariance selection in Gaussian graphical model

- Let $X = (X^{(1)}, ..., X^{(p)}) \sim N_p(\mu, \Sigma)$ with μ denoting the unknown mean and Σ denoting the nonsingular covariance matrix. We wish to estimate the concentration matrix $C = \Sigma^{-1}$.
- The zero entries c_{ij} in the concentration matrix indicates the conditional independence between the two random variables $X^{(i)}$ and $X^{(j)}$ given all other variables (Dempster, 1972, Whittaker, 1990, Lauritzen, 1996).
- The Gaussian random vector X can be represented by an undirected graph $G = (V, E)$, where V contains p vertices corresponding to the p coordinates and the edges $E = (e_{ij})_{1 \leq i < j \leq p}$ represent the conditional dependency relationships between variables $X^{(i)}$ and $X^{(j)}$.

Modified BIC for graphical model

Under this high-dimensional setup, the penalized likelihood estimation of covariance matrix has been investigated by Rothman et al. (2008) and Lam and Fan (2009).
We propose to modify the BIC with an extra penalty term of $4 \log p_n$ on the dimension of the covariance matrix, while the log-likelihood term is evaluated directly at the penalized estimator. Given a λ, the associated modified BIC criterion is defined as:

$$BIC_\lambda = -n \log |\hat{C}_\lambda| + n\mathrm{tr}(\hat{C}_\lambda \bar{A}) + \{\log n + 4 \log p_n\} \sum_{1 \le i < j \le p} I(\hat{c}_{ij,\lambda} \neq 0$$

Modified BIC for graphical model

We further assume that d_T is bounded by a finite constant Q and $(p_n/n)(\log p_n)^k = O(1)$, for some $k > 1$.

Theorem

Under the regularity conditions, $Pr(G_{\hat{\lambda}_{BIC}} = G_T) \to 1$, where $\hat{\lambda}_{BIC}$, which may not be unique, is the tuning parameter that minimizes the modified BIC criterion with the SCAD penalty.

simulation

Table: Results for Graphical Model with p=250 and N=500.

	LASSO				*SCAD*			
	model 1		model 2		model 1		model 2	
	bic	cv	bic	cv	bic	cv	bic	cv
fp	16.30	566.24	10.22	952.57	2.94	1981.43	7.99	90.46
	(5.92)	(28.12)	(5.40)	(68.63)	(1.72)	(47.17)	(4.22)	(13.37)
fn	0.12	0.00	110.01	23.48	0.36	0.00	112.80	76.93
	(0.38)	(0.00)	(8.49)	(3.95)	(0.72)	(0.00)	(7.58)	(4.82)
tp	98.88	99.00	79.99	166.52	223.64	224.00	202.20	238.07
	(0.38)	(0.00)	(8.49)	(3.95)	(0.72)	(0.00)	(7.58)	(4.82)
tn	31010	30460	30925	29982	31023	29045	30927	30845
	(5.92)	(28.12)	(5.40)	(68.63)	(1.72)	(47.17)	(4.22)	(13.37)
spec	1.00	0.98	1.00	0.97	1.00	0.94	1.00	1.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
sens	1.00	1.00	0.42	0.88	1.00	1.00	0.64	0.76
	(0.00)	(0.00)	(0.04)	(0.02)	(0.00)	(0.00)	(0.02)	(0.02)
mcc	0.93	0.38	0.61	0.35	0.99	0.31	0.78	0.74
	(0.02)	(0.01)	(0.03)	(0.01)	(0.00)	(0.00)	(0.01)	(0.02)
fdr	0.14	0.85	0.11	0.85	0.01	0.90	0.04	0.27
	(0.04)	(0.01)	(0.05)	(0.01)	(0.01)	(0.00)	(0.02)	(0.03)
psr	1.00	1.00	0.42	0.88	1.00	1.00	0.64	0.76
	(0.00)	(0.00)	(0.04)	(0.02)	(0.00)	(0.00)	(0.02)	(0.02)

SCAD:the SCAD penalty; LASSO: the L_1 penalty.

Conclusion

- Information Criterion can be constructed based on composite likelihood or other pseudo-likelihood when full likelihood is hard to compute
- Extra penalty needed on the dimensionality of the model space under $P \to \infty$ case.
- Can be used to select variables in mean structure or select sparse covariance or inverse covariance structure
- The likelihood term can be evaluated at the penalized likelihood estimator so that the resulting BIC can be directly used to select the tuning parameters for penalized likelihood estimation