Efficiency of bivariate composite likelihood

Harry Joe
Department of Statistics
University of British Columbia

Abstract:

Composite likelihood, including pairwise likelihood, procedures can be used for multivariate models with regression/univariate and dependence parameters, when the joint multivariate probability or density is computationally too difficult. We have studied composite likelihood methods for some models for clustered/longitudinal data and times series data. In some cases, the asymptotic efficiency of composite likelihood estimation can be studied via the theory of estimating equations. For clustered data with varying cluster sizes, for several models we study asymptotic relative efficiencies for various weighted pairwise likelihoods, with weight being a function of cluster size. For longitudinal data, we also consider weighted pairwise likelihoods with weights that can depend on lag. Based on our analyses, we can give some recommendations on the choice of weights to achieve better efficiency. We also report on some results for bivariate composite likelihood applied to a stochastic volatility model for financial time series.

Outline

1. Composite likelihood as part of inference based on low-dimensional margins.
2. Examples where I have used composite likelihood
3. Choice of weights for bivariate composite likelihood for clustered data : cluster size d_i, weight $w_i \propto w(d_i)$
4. Choice of margins for models for longitudinal and time series data

bivariate composite log-likelihood (BCL); same as pairwise log-likelihood but P in PL can be pairwise, pseudo, penalized or partial
also have extension to trivariate composite likelihood (TCL)
Goals: To show relative efficiency of BCL and not computational details; to show when BCL can have good or poor performance.

My research in composite likelihood is part of more general research for multivariate inference based on low-dimensional margins.

There are several methods for inference based on low-dimensional margins, and they are called limited information methods in psychometrics.

Used when

- high-dimensional probability/density numerically too time-consuming to compute
- data sparse in high dimensions but not sparse in low-dimensional margins.

Examples:

1. Inference function for margins (IFM): different parameters for different univariate margins; fit separate univariate models, then consider several different mult models with the given univariate margins. IFM is useful if the joint log-likelihood is not simple and the total number of univariate plus multivariate parameters is large (say, > 20).

2. Composite likelihood: Familial data with several different familial correlations, and common univariate regression parameters for different members of family; multivariate probit model (binary response), multivariate normal copula/mixture models (count response, survival data).

3. Item response data: m items (say $m \geq 20$), ordinal response with K categories per item; m-dimensional table with K^m categories is sparse, but want inferences for model, including GOF.

Item response data: m items, K categories per item, n=sample size, n/K^m small; e.g., $K=5$, $m=10$.

Logit-normit model or graded logistic with one latent trait: $\boldsymbol{\theta} = (\boldsymbol{\alpha}', \boldsymbol{\beta}')'$, $\boldsymbol{\alpha} = (\alpha_{i1}, \ldots, \alpha_{i,K-1} : i = 1, \ldots, m)'$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_m)'$,

$$\Pr\left[\cap_{i=1}^m \{Y_i = y_i\}\right] = \int_{-\infty}^{\infty} \prod_{i=1}^m \Pr(Y_i = y_i \mid \eta)\, \phi(\eta) d\eta, \quad y_i \in \{0, \ldots, K-1\}$$

$$\Pr(Y_i = j \mid \eta) = \begin{cases} 1 - G(\alpha_{i,1} + \beta_i \eta) & \text{if } j = 0 \\ G(\alpha_{i,j} + \beta_i \eta) - G(\alpha_{i,j+1} + \beta_i \eta) & \text{if } 0 < j < K-1 \\ G(\alpha_{i,K-1} + \beta_i \eta) & \text{if } j = K-1 \end{cases}$$

where $\phi(\cdot)$ is the standard normal density, G is logistic cdf.
Rasch submodel has $\boldsymbol{\beta} = \beta \mathbf{1}$ (constant slope).

For our family of quadratic form GOF statistics (Maydeu & Joe 2005 in JASA, 2006 in Psychometrika), either MLE or bivariate composite log-likelihood (BCL) or any $\sqrt{n}$-consistent estimator can be used, and theory is similar.
For estimation, expected Fisher information cannot be computed when K^m is too large, but asymptotic covariance matrix of the BCL estimator can be computed.
Efficiency of BCL slowly decreases as number of items m increases, but above 0.98 for less than 10 items and 2–5 categories per item.

Familial data: neurofibromatosis, many phenotypes: binary (presence of symptom) and count (#tumours); one covariate is genotype = type of mutation
Y binary, $\mathbf{x}$ covariates, Z latent variable, $Y = I(Z \le \alpha + \boldsymbol{\beta}^T \mathbf{x})$.
Probit: $Z \sim$ N(0,1), $F_Z(z) = \Phi(z)$.
In genetics, latent var. = liability, threshold=τ, mean liability depends on covariates.
If $\pi(\mathbf{x}) = \Pr(Z \le \alpha + \boldsymbol{\beta}'\mathbf{x})$, then $F_Z^{-1}(\pi(\mathbf{x})) = \alpha + \boldsymbol{\beta}'\mathbf{x}$.

Probit model for multivariate binary: $(Y_1, \ldots, Y_d)$ binary response vector for a family of size d, $\mathbf{x}_1, \ldots, \mathbf{x}_d$ covariate vectors, $Y_j = I(Z_j \le \alpha + \boldsymbol{\beta}'\mathbf{x}_j)$.

$(Z_1, \ldots, Z_d) \sim N(\mathbf{0}, R)$: where correlations in $R = R(\rho_{ss}, \rho_{po}, \rho_2, \rho_3)$ depend on the relationships of the family members.

Model for count data

For a count response, a useful model with a general dependence structure is the Poisson-lognormal mixture model.

family i, member j within family, count Y_j
$Y_j | \Lambda_j = \lambda_j \sim$ Pois(λ_j) and $\log \Lambda_j \sim$ N(μ_j, σ^2), where $\mu_j = \alpha + \boldsymbol{\beta}' x_j$,
$(\log \Lambda_1, \ldots, \log \Lambda_d) \sim N_d(\boldsymbol{\mu}, \Sigma)$ with correlation parameters
$\rho_{ss}, \rho_{po}, \rho_2, \rho_3$.

[sib-sib, parent-offspring, degree 2, degree 3]

With these multiv models (probit or Poisson random effects), $f_{Y_1,\ldots,Y_d}$ involves a d-dimensional numerical integral, with computation complexity increasing exponentially with d.

For a continuous variable Y (e.g., onset time) that can be right-censored, use of the multivariate normal distribution means that the dimension of the numerical integral is equal to the number of censored observations.

There are parameters common to different margins, e.g., common regression parameters for different univariate margins, and common dependence parameters for different bivariate margins.

Because family size (e.g., multi-generation families) can be large and is varying, composition likelihood methods based on sum of log-likelihoods of univariate/bivariate margins were considered.

[Zhao and Joe, 2005, CJS]
CL1 method: estimate univariate parameters $\boldsymbol{\theta}$ by maximizing :

$$\Psi_1(\boldsymbol{\theta}) = \sum_i \sum_j l_1(y_{ij}; \boldsymbol{\theta}, \mathbf{x}_{ij}),$$

where $l_1(y_{ij}; \boldsymbol{\theta}, \mathbf{x}_{ij})$ is the univariate marginal log-likelihood of Y_{ij}. Then estimate dependence parameters $\boldsymbol{\delta}$ by maximizing

$$\Psi_2(\hat{\boldsymbol{\theta}}, \boldsymbol{\delta}) = \sum_i \sum_{j>j'} l_2(y_{ij}, y_{ij'}; \hat{\boldsymbol{\theta}}, \boldsymbol{\delta}; \mathbf{x}_{ij}, \mathbf{x}_{ij'}),$$

wrt $\boldsymbol{\delta}$, where $l_2(y_{ij}, y_{ij'}; \boldsymbol{\theta}, \boldsymbol{\delta}; \mathbf{x}_{ij}, \mathbf{x}_{ij'})$ is the bivariate log likelihood of $(Y_{ij}, Y_{ij'})$ and $\hat{\boldsymbol{\theta}}$ is the estimate of $\boldsymbol{\theta}$ obtained in the first step.

CL2 method: estimate $\boldsymbol{\theta}$ and $\boldsymbol{\delta}$ in one step by maximizing

$$\Psi_2^*(\boldsymbol{\theta}, \boldsymbol{\delta}) = \sum_i w_i \sum_{j>j'} l_2(y_{ij}, y_{ij'}; \boldsymbol{\theta}, \boldsymbol{\delta}; \mathbf{x}_{ij}, \mathbf{x}_{ij'}),$$

where w_i is a weight depending on the family size. Our choice of weight was $w_i = (d_i - 1)^{-1}$ [explanations later]; this choice also used by others. $w_i = 1$ puts too much weight on large families.

For the estimated covariance matrix for CL1 or CL2, one approach is the jackknife if the derivatives for the Godambe matrix are difficult to obtain.

Theory and Efficiency results - familial data: Zhao (2004) PhD thesis, Zhao, Y and Joe, H (2005). Composite likelihood estimation in multivariate data analysis, *Canad J Statist*, 33, 335–356.

Analysis via theory of estimating equations; theoretical limits for independent and perfect dependence, and numerical calculations for intermediate dependence.
Relative efficiency : $A\mathrm{Var}\,(\hat{\theta}_{MLE,k})/A\mathrm{Var}\,(\hat{\theta}_{CL,k})$

Performance of the two methods is reasonable, except that when the dependence is strong, the first approach is inefficient for the regression parameters. The second approach is generally better for the regression parameters, but may be less efficient for the dependence parameters when the dependence is weak.

[Joe & Lee, 2008] My collaborator Y. Lee thought that choice of weights should be considered more carefully in clustered data with varying cluster size.

Data $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{id_i})$, $i = 1, \ldots, n$, covariates $(\mathbf{x}_{i1}, \ldots, \mathbf{x}_{id_i})$.

$$L_{BCL}(\boldsymbol{\theta}) = \sum_i \sum_{1 \le j < k \le d_i} w_{i,jk} L_{jk}(\boldsymbol{\theta}),$$

where $w_{i,jk}$ = weight that depends on d_i for clustered data,
$w_{i,jk}$ can depend on lag for longitudinal data.
Set of estimating equations:

$$\mathbf{g}(\boldsymbol{\theta}) = \partial L_{BCL}/\partial \boldsymbol{\theta} \tag{1}$$

Under regularity conditions, $\tilde{\boldsymbol{\theta}}_{BCL}$ is the solution of (1), and the asymptotic cov matrix of $n^{1/2}(\tilde{\boldsymbol{\theta}}_{BCL} - \boldsymbol{\theta})$ has form
$\mathbf{V} = \mathbf{D}^{-1}\mathbf{M}(\mathbf{D}')^{-1}$ where $\mathbf{D} = -\mathrm{E}\,[\partial \mathbf{g}/\partial \boldsymbol{\theta}]$, $\mathbf{M} = \mathrm{Cov}\,(\mathbf{g})$.

Cases where BCL is same as log-likelihood

1. Observations are independent within each cluster: $f_{\mathbf{Y}_i} = \prod_{j=1}^{d_i} f_{Y_{ij}}$, choice $w_{i,jk} = 1/(d_i - 1)$ results in log-likelihood.
2. Discrete observations such that perfect dependence (Fréchet upper bound) holds: for all clusters $y_{i1} = \cdots = y_{id_i}$, $f_{\mathbf{Y}_i}(\mathbf{y}_i) = f_{Y_{i1}}(y_{i1})$ for $\mathbf{y}_i = y_{i1}\mathbf{1}$, and $f_{Y_{ij},Y_{ik}}(y_{ij}, y_{ik}) = f_{Y_{i1}}(y_{i1})$ for $y_{ij} = y_{ik} = y_{i1}$; choice $w_{i,jk} = 1/[d_i(d_i - 1)]$ results in log-likelihood.

Cases 1 and 2 mean that the optimal weights should depend on the amount of dependence within clusters.

---

3. Longitudinal data based on AR(1) time series, log-lik is:

$$\sum_i \Big[-\sum_{j=2}^{d_i-1} \log f_{Y_{ij}}(y_{ij};\boldsymbol{\alpha}) + \sum_{j=2}^{d_i} f_{Y_{i,j-1},Y_{ij}}(y_{i,j-1}, y_{ij};\boldsymbol{\theta}) \Big],$$

where $\boldsymbol{\alpha}$=subset of univariate parameters. If $\boldsymbol{\alpha}$ assumed known, then BCL=log-likelihood when $w_{i,jk} = 1$ for $k = j + 1$ and 0 otherwise.

Case 3 suggests weights that depend on lag for longitudinal data; see also Varin & Vidoni (2006, CSDA)

Models for which efficiency calculations can be made: multivariate exchangeable normal (one-way random effects with varying cluster sizes) and probit; these provide insight on efficiency.

For the exchangeable d-variate normal distribution, the mean vector and covariance matrix are respectively

$$\mu\mathbf{1}_d \text{ and } \Sigma_d = \eta^2\mathbf{R}(\rho),$$

where $\mathbf{R}(\rho) = [(1 - \rho)\mathbf{I}_d + \rho\mathbf{J}_d]$, $\mathbf{I}_d$ is the identity matrix of order p and $\mathbf{J}_d$ is the $d \times d$ matrix of 1s. Univariate parameters are mean μ , variance η^2, and dependence parameter is ρ.
Exchangeable probit:

$$Y_j = I(Z_j \le \mu), \quad j = 1, \ldots, d, \qquad (Z_1, \ldots, Z_d)' \sim N(\mathbf{0}, \mathbf{R}(\rho)).$$

Optimal weight w_i as a function of d_i depends on parameter (μ, η^2 or ρ).

Varying cluster size; est'n of μ with ρ known; exch normal

For the BCL, if ρ were known, optimal w_i is

$$w_i = (d_i - 1)^{-1}[1 + (d_i - 1)\rho]^{-1}. \tag{2}$$

Consider some weights that do not depend on ρ:

(a) $w_i = 1$ (unweighted);

(b) $w_i = (d_i - 1)^{-1}$ (matches loglik for independence);

(c) $w_i = (d_i - 1)^{-1}[1 + \frac{1}{2}(d_i - 1)]^{-1}$ (from subst $\rho = \frac{1}{2}$ in (2));

(d) $w_i = (d_i - 1)^{-1}d_i^{-1}$ (matches loglik for perfect dependence)

Varying cluster size; est'n of ρ with μ, η^2 known; exch normal

For $\rho = 0, 0.5, 1$, optimal weight is:

$$w_i \propto \begin{cases} 1, & \rho = 0; \\ [0.0625d_i^2 + 1.0625d_i + 0.4375]^{-1}, & \rho = 0.5; \\ d_i^{-1}, & \rho = 1. \end{cases}$$

For 2 cases of moderate to strong dependence, inverse weight is close to linear for small d_i.
Varying cluster size, est'n of η^2 with ρ, μ known; exch normal
Optimal choice of $w_i \propto t_i = d_i^2\rho^2 + d_i(1 + 2\rho - 3\rho^2) - (1 + 2\rho - 3\rho^2)$.

$$w_i \propto \begin{cases} (d_i - 1)^{-1}, & \rho = 0; \\ [0.25d_i^2 + 1.25d_i - 1.25]^{-1}, & \rho = 0.5; \\ d_i^{-2}, & \rho = 1. \end{cases}$$

In the middle case of moderate dependence, the inverse weight for small d_i is close to $(d_i - 1)[1 + \frac{1}{2}(d_i - 1)]$ [(c) in previous slide].
The other two cases of inverse linear and quadratic weights occurred previously.

Exchangeable probit: estimating two parameters simultaneously

Some REs of the BCL estimates of the two parameters for four sets of weights, labeled (a)–(d) previously, with different distributions of cluster sizes. The 3 settings for ρ represent weak, moderate and strong correlation. The patterns are similar for different μ.

	wt :	(a) $w_i = 1$		(b) $(d_i - 1)^{-1}$		(c) interm		(d) $(d_i - 1)^{-1}d^{-1}$	
ρ	mixture	REμ	REρ	REμ	REρ	REμ	REρ	REμ	REρ
0.2	0.5 for $d = 3, 4$	.949	.986	.998	.978	.995	.941	.990	.927
	0.5 for $d = 3, 5$	.879	.969	.994	.956	.985	.853	.970	.818
	0.5 for $d = 3, 6$	.830	.953	.988	.939	.974	.770	.948	.720
	0.5 for $d = 3, 7$	.800	.939	.981	.927	.962	.700	.928	.641
	0.5 for $d = 3, 8$	.780	.927	.974	.920	.951	.641	.909	.577
	0.2 for $d = 3 \ldots 7$	.859	.951	.990	.956	.981	.827	.963	.785
0.6	0.5 for $d = 3{,}4$	.921	.954	.988	.983	.997	.967	.995	.957
	0.5 for $d = 3, 5$	.821	.895	.967	.971	.995	.929	.990	.907
	0.5 for $d = 3, 6$	.751	.844	.942	.958	.993	.894	.985	.860
	0.5 for $d = 3, 7$	.705	.806	.918	.943	.992	.863	.980	.820
	0.5 for $d = 3, 8$	.676	.777	.895	.929	.991	.836	.976	.786
	0.2 for $d = 3 \ldots 7$	.791	.850	.953	.951	.993	.911	.987	.886
0.9	0.5 for $d = 3, 4$	.899	.940	.974	.985	.989	.980	.989	.974
	0.5 for $d = 3, 5$	.781	.857	.939	.966	.985	.956	.985	.941
	0.5 for $d = 3, 6$	.702	.791	.902	.944	.982	.933	.982	.910
	0.5 for $d = 3, 7$	.651	.743	.867	.919	.979	.912	.979	.882
	0.5 for $d = 3, 8$	.617	.709	.836	.895	.976	.893	.976	.859
	0.2 for $d = 3 \ldots 7$	.747	.804	.919	.936	.979	.934	.979	.918

Exchangeable probit: representative of a model where BCL is needed for computational reasons

1. The REs of the BCL estimates decrease with larger cluster sizes and more variability in cluster sizes.
2. $w_i = 1$ weights not good, esp. for univariate parameter.
3. Best w_i depends on parameter and strength of dependence.

   (i) Weak dependence: $w_i = 1/(d_i - 1)$ best for μ, $w_i = 1$ best for ρ, $w_i = 1/(d_i - 1)$ best overall.

   (ii) Moderate dep: $w_i = 1/[(d_i - 1)(1 + 0.5(d_i - 1))]$ best for μ, $w_i = 1/(d_i - 1)$ best for ρ, $w_i = 1/(d_i - 1)$ or $w_i = 1/[(d_i - 1)(1 + 0.5(d_i - 1))]$ best.

   (iii) Strong dependence: $w_i = 1/[(d_i - 1)d_i]$ best for μ, $w_i = 1/(d_i - 1)$ best for ρ, $w_i = 1/[(d_i - 1)(1 + 0.5(d_i - 1))]$ best overall.

Longitudinal data

AR(1) normal

In order to see some patterns, we study the AR(1) model $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{id})' \sim N(\mu \mathbf{1}_d, \eta^2 \mathbf{R}(\rho))$ where $\mathbf{R} = \mathbf{R}(\rho) = (\rho^{|j-k|})_{1 \le j,k \le d}$ and $-1 < \rho < 1$. Some analytic results are possible for the estimation of one of the parameters assuming the other two to be known.

Let L_0 denote the log-likelihood, L_1 denote the BCL with all pairs, L_2 denote the BCL with only adjacent pairs, and L_3 denote the BCL with adjacent pairs but assuming the first and the last observations to be adjacent. A justification is that, by adding the $(1,d)$ bivariate margin to the adjacent bivariate margins, the resulting BCL becomes twice the log-likelihood at the boundary case of independence ($\rho = 0$), so that maybe it can be expected to do best under weak dependence.

Patterns of the relative efficiency (RE) functions

1. For estimation of μ, generally best is BCL with adjacent pairs including $(1,d)$.
2. For estimation of η^2, for most (d,ρ) pairs, best is BCL with adjacent pairs including $(1,d)$. But RE goes to 0 as $\rho \to \pm 1$.
3. The smallest REs for $\widehat{\mu}_2$ or $\widehat{\mu}_3$ [adjacent pairs] are bounded below by 0.877, but that for $\widehat{\eta}_1^2$ [all pairs] can get much smaller for larger $|\rho|$ and d. In a minimax sense, it is better to use all adjacent bivariate margins (plus $(1,d)$ margin for small d) than to use all bivariate margins.

AR(1) binary probit

AR(1) probit model without covariates has 2 parameters: $Y_j = I(Z_j \le \mu)$, $j = 1, \ldots, d$, $(Z_1, \ldots, Z_d)' \sim N(\mathbf{0}, \mathbf{R}(\rho))$, $\mathbf{R}(\rho) = (\rho^{|j-k|})_{1 \le j,k \le d}$. With data $(y_{i1}, \ldots, y_{id})'$, $i = 1, \ldots, n$, consider weighted BCL :

$$L_{BCL} = \sum_{\ell=1}^{d-1} w_\ell \sum_{j=1}^{d-\ell} L_{j,j+\ell}$$

where the weight of the (j,k) bivariate margin depends on the lag $\ell = k - j$, and

$$L_{jk} = \sum_{s=0}^{1} \sum_{t=0}^{1} n_{st}^{(jk)} \log p_{jk}(s,t;\mu,\rho),$$

$p_{jk}(s,t;\mu,\rho) = \Pr(Y_j = s, Y_k = t; \mu, \rho)$. Intuitively, we want to mainly use bivariate margins with lag 1 in order to reduce the amount of computation for larger d.

From actual efficiency calculations, choice of $w_1 = w_{d-1} = 1$ and other $w_\ell = 0$ is good.

As a preliminary analysis on the choice of weights, for small values of d, we did a regression analysis of $\log \Pr(Y_1 = y_1, \ldots, Y_d = y_d; \mu, \rho)$ on $\log \Pr(Y_j = y_j, Y_k = y_k; \mu, \rho)$ (all $1 \le j \le k \le d$) with 'data' collected from inputs with different $(y_1, \ldots, y_d, \mu, \rho)$.

For $d = 3, 4, 5, 6$, the pattern is as follows:

(i) the largest regression coefficients were for the adjacent pairs $\log \Pr(Y_1 = y_1, Y_2 = y_2)$ and $\log \Pr(Y_{d-1} = y_{d-1}, Y_d = y_d)$,

(ii) the second largest coefficients were for the adjacent pairs $\log \Pr(Y_2 = y_2, Y_3 = y_3), \ldots, \log \Pr(Y_{d-2} = y_{d-2}, Y_{d-1} = y_{d-1})$

(iii) third largest coefficient was for $\log \Pr(Y_1 = y_1, Y_d = y_d)$. The regression R^2 did not change much when other non-adjacent pairs were deleted. When log of univariate marginal probabilities $\log \Pr(Y_j = y_j)$ were included, the additional regression coefficients were close to zero.

Table 4: AR(1) probit

d	μ	ρ	$w_1,\ldots,w_{d-1}$	RE($\widehat{\mu}_w$)	RE($\widehat{\rho}_w$)
3	-0.52	0.5	1.0,1.0	0.996	0.965
3	-0.52	0.5	1.0,0.0	0.912	0.946
3	-0.52	0.5	0.0,1.0	0.877	0.357
3	-0.52	0.5	1.0,0.8	0.990	0.983
4	-0.52	0.5	1.0,1.0,1.0	0.991	0.922
4	-0.52	0.5	1.0,0.0,0.0	0.899	0.925
4	-0.52	0.5	1.0,0.0,1.0	0.993	0.933
4	-0.52	0.5	0.0,1.0,0.0	0.987	0.436
4	-0.52	0.5	0.0,0.0,1.0	0.764	0.124
4	-0.52	0.5	1.0,1.0,0.0	0.945	0.952
4	-0.52	0.5	1.0,0.5,0.0	0.930	0.986
5	-0.52	0.5	1.0,1.0,1.0,1.0	0.988	0.887
5	-0.52	0.5	1.0,0.0,0.0,0.0	0.897	0.914
5	-0.52	0.5	1.0,0.0,0.0,1.0	0.991	0.919
5	-0.52	0.5	0.0,1.0,0.0,0.0	0.911	0.474
5	-0.52	0.5	0.0,0.0,1.0,0.0	0.933	0.169
5	-0.52	0.5	1.0,1.0,0.0,0.0	0.912	0.948
5	-0.52	0.5	1.0,0.0,1.0,0.0	0.956	0.920
5	-0.52	0.5	1.0,1.0,1.0,0.0	0.957	0.907
5	-0.52	0.5	1.0,1.0,0.0,1.0	0.971	0.941
5	-0.52	0.5	1.0,0.0,1.0,1.0	0.991	0.889
5	-0.52	0.5	1.0,0.5,0.0,0.0	0.908	0.981
6	-0.52	0.5	1.0,1.0,1.0,1.0,1.0	0.985	0.863
6	-0.52	0.5	1.0,0.0,0.0,0.0,0.0	0.900	0.908
6	-0.52	0.5	1.0,0.0,0.0,0.0,1.0	0.990	0.911
6	-0.52	0.5	1.0,1.0,0.0,0.0,0.0	0.900	0.947

Single time series with a latent AR(1) process, consider the AR stochastic volatilty ARSV(1) model:

$$Y_t = \sigma_t \epsilon_t, \quad \log \sigma_t^2 = \alpha + \rho \log \sigma_{t-1}^2 + V_t$$

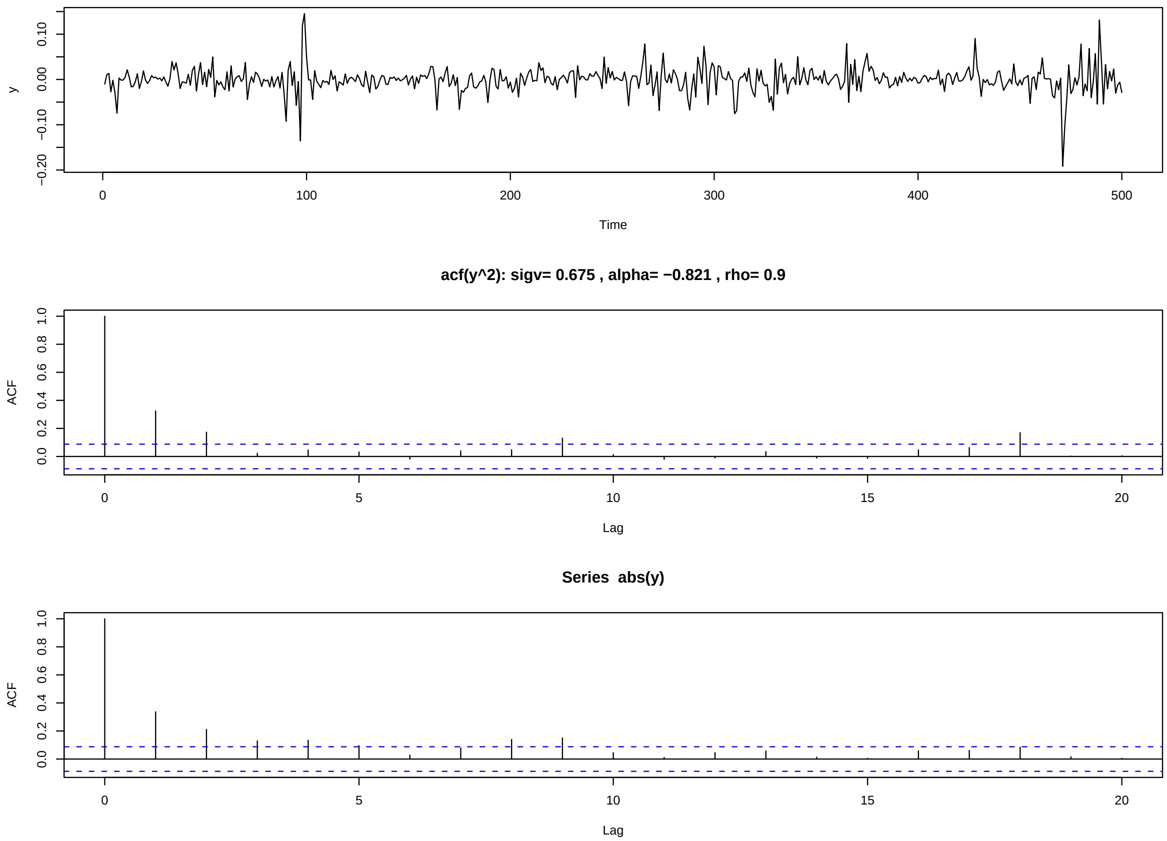
with $-1 < \rho < 1$, where ϵ_t are iid N(0,1) and V_t are iid $\mathrm{N}(0, \sigma_V^2)$, $\{\epsilon_t\}$ and $\{V_t\}$ are mutually independent, α is a constant. $\xi = \alpha/(1-\rho), \omega = \sigma_v/\sqrt{1-\rho^2}$ are the mean and SD of $\log \sigma_t^2$.

Given observations $y_1, \ldots, y_n$, the joint density is

$$f_{Y_1,\ldots,Y_n}(y_1,\ldots,y_n) = \int \Big\{ \prod_{j=1}^{n} s_j^{-1} f_\epsilon(y_j/s_j) \Big\} f_{\sigma_1,\ldots,\sigma_n}(s_1,\ldots,s_n)\, ds_1 \cdots ds_n.$$

$f_{\sigma_1,\ldots,\sigma_n}$ involves a density for the AR(1) process $\{\log \sigma_t^2\}$. The dimension of the integrand increases with n.

Bivariate and trivariate marginal densities can easily be computed via Gauss-Hermite quadrature.


acf(y^2): sigv= 0.675 , alpha= −0.821 , rho= 0.9

Series abs(y)

The ARSV is an alternative to ARCH/GARCH models for financial time series, but computationally it is harder to work with.

Some significant autocorrelations in $|y_t|$ for $0.6 < \rho < 1$ and $\sigma_L(\rho) < \sigma_V < \sigma_U(\rho)$ (not too small/large), where $\sigma_L(\rho)$ is decreasing in ρ and $\sigma_U(\rho)$ is increasing in ρ.

For financial time series data, often estimated ρ exceeds 0.9.

From Monte Carlo simulations, some results on relative efficiency of BCL compared with ML (simulated likelihood in Ox) and with quasi-ML (quasi-likelihood is based on using MVN density for $(\log Y_1^2, \ldots, \log Y_n^2)$. Have more efficient implementation of quasi-ML than Kalman filter.

Summaries for $\rho = .7$, $\rho = .9$, $\rho = .95$ are given in following tables.

Table ARSV1: $\rho = 0.9$, $\sigma_V = 0.675$, $\alpha = -0.821$ (used in Fridman-Harris and Sandmann-Koopman etc). $\xi = \alpha/(1-\rho)$, $\omega = \sigma_V/(1-\rho^2)^{1/2}$. root MSE of parameter estimates for different methods. Efficiency is average relative root MSE (over parameters) with respect to ML. 100 simulations, sample size 500. RE depends on parametrization

parameter	ML/Ox	BCL$_1$	BCL$_2$	BCL$_{LB}$	QML
σ_V	.086	.21	.17	.18	.18
ρ	.037	.089	.070	.062	.074
α	.31	.74	.58	.52	.61
ξ	.27	.32	.32	.32	.34
ω	.17	.17	.18	.18	.19
avg eff	1	.62	.67	.69	.63

BCL$_1$ based on margins $(i, i+1)$; BCL$_2$ based on margins $(i, i+1)$ and $(i, i+2)$; BCL$_{LB}$ based on lower bound 0.8 for $\widehat{\rho}$.
Best is TCL with $(i, i+1, i+2)$ and $(i, i+1, i+3)$.

Table ARSV2: $\rho = 0.95$, $\sigma_V = 0.260$, $\alpha = -0.368$ (used in Fridman-Harris and Sandmann-Koopman etc). $\xi = \alpha/(1-\rho)$, $\omega = \sigma_V/(1-\rho^2)^{1/2}$. root MSE of parameter estimates for different methods. Efficiency is average relative root MSE (over parameters) with respect to ML. 100 simulations, sample size 500.

parameter	ML/Ox	BCL$_1$	BCL$_2$	BCL$_{LB}$	QML
σ_V	.060	.21	.17	.13	.22
ρ	.042	.22	.17	.084	.21
α	.31	1.6	1.2	.62	1.7
ξ	.20	.25	.25	.25	.27
ω	.14	.16	.16	.16	.20
avg eff	1	.47	.51	.63	.42

BCL: negative bias for α, ρ.
BCL efficiency gets worse as ρ gets closer to 1 (same also for QML) mainly due to distributions for $\widehat{\rho}$ and $\hat{\alpha}$ that are very negatively skewed but clustered near their true values.

Table ARSV3: $\rho = 0.7$, $\sigma_V = 1.8$, $\alpha = 0.1$ $\xi = \alpha/(1-\rho)$, $\omega = \sigma_V/(1-\rho^2)^{1/2}$. root MSE of parameter estimates for different methods. Efficiency is average relative root MSE (over parameters) with respect to ML. 100 simulations, sample size 500.

parameter	ML/Ox	BCL$_1$	BCL$_2$	QML
σ_V	.29	.18	.14	.20
ρ	.063	.062	.048	.065
α	.071	.10	.10	.10
ξ	.27	.32	.32	.31
ω	.27	.16	.15	.18
avg eff	1	1.2	1.3	1.1

ML with negative bias for σ_V and ω for smaller ρ.

Conclusions

1. Composition likelihood methods are easier to implement than maximum likelihood when multivariate density is difficult to compute; just need software with numerical optimization and low-dimensional integration.

2. Bivariate composite likelihood generally has good efficiency except sometimes when dependence is strong. Worst case for multivariate discrete not as bad as worse case for multivariate continuous.

3. For clustered data with varying cluster size d_i, the optimal weight depends on the parameter and the strength of association.

4. For clustered data, $w_i \propto (d_i-1)^{-1}[1+\frac{1}{2}(d_i-1)]^{-1}$ is generally good.

5. For longitudinal data/time series, bivariate margins with lag 1 could be considered as a starting point. Adding the $(1, d)$ margin [adjacent with wraparound for indices] helps for small cluster size d. Adding $(i, i+2)$ margins with lag 2 might also improve efficiency.