Average information splitting and multi-frontal solvers for heterogeneous high-throughput data analysis

Shengxin.Zhu@xjtlu.edu.cn
Xi'an Jiaotong-Liverpool University, Suzhou, China

A special application dedicated to Iain Duff's 70th birthday

▷ Background
Motivation
LMMs & REML
Problems
Challenges
Methods
Results
Discussion

Background

Motivation

▷ Background
▷ Motivation
LMMs & REML
Problems
Challenges
Methods
Results
Discussion

Analyze multi-sources unbalanced/heterogenous data.

- Astrophysics with discrete sources,
  - J.-P. Roques et al INTEGRAL SPI observation of the galactic central radian: Contribution of discrete sources and implication for the diffuse emission 1, The Astrophysical Journal, 635 (2005)
  - Amestoy, Duff et al On Computing Inverse Entries of a Sparse Matrix in an Out-of-Core Environment, SIAM J. SCi., 2012

Motivation

▷ Background
▷ Motivation
LMMs & REML
Problems
Challenges
Methods
Results
Discussion

Analyze multi-sources unbalanced/heterogenous data.

- Astrophysics with discrete sources,
  - Amestoy, Duff et al On Computing Inverse Entries of a Sparse Matrix in an Out-of-Core Environment, SIAM J. SCi., 2012
- Animal/plant breeding, clinic trial, genome-wide association
  - Welham, Zhu, and Wathen, Big data, fast models: faster calculation of models from high-throughput biological data sets, Knowledge Transfer Project Reprot IP12-009, 2013, Smith Institute, Oxford and VSNi (was $\in$ NAG)

Motivation

▷ Background
▷ Motivation
LMMs & REML Problems
Challenges
Methods
Results
Discussion

Analyze multi-sources unbalanced/heterogenous data.

- Astrophysics with discrete sources,
  - Amestoy, Duff et al On Computing Inverse Entries of a Sparse Matrix in an Out-of-Core Environment, SIAM J. SCi., 2012
- Animal/plant breeding, clinic trial, genome-wide association
  - Welham, Zhu, and Wathen, Big data, fast models: faster calculation of models from high-throughput biological data sets, Knowledge Transfer Project Reprot IP12-009, 2013, Smith Institute, Oxford and VSNi (was $\in$ NAG)
  - Lippert et al, Fast linear mixed models for genome-wide association study, *Nature Methods*, 8(10) , 2011
  - Listgarten et al, Improved linear mixed models for genome-wide association studies, *Nuture Methods*, 9(6),2012
  - Zhang et al, Mixed linear model approach adapted for genome-wide association studies, *Nature Genetics*, 42(4), 2010.
  - Zhou et al, Mixed linear model approach for genome-wide association studies, *Nature Genetics*,44(7) 2012
  - Bolker et al, Generalized linear mixed models: a practical guide for ecology and evolution, *Trends in Ecology and Evolution*,24(3), 2008

Motivation

▷ Background
▷ Motivation
LMMs & REML Problems
Challenges
Methods
Results
Discussion

Analyze multi-sources unbalanced/heterogenous data.

- Astrophysics with discrete sources,
  - Amestoy, Duff et al On Computing Inverse Entries of a Sparse Matrix in an Out-of-Core Environment, SIAM J. SCi., 2012
- Animal/plant breeding, clinic trial, genome-wide association
  - Welham, Zhu, and Wathen, Big data, fast models: faster calculation of models from high-throughput biological data sets, Knowledge Transfer Project Reprot IP12-009, 2013, Smith Institute, Oxford and VSNi (was $\in$ NAG)
- Multi-run measurements, multi-sources data analysis.
- Driver-assistance systems, automated driving
- 3D imaging processing(3D SPECT,3D PET)
- Cyber-physical systems $\cdots$

Common features

- Unbalanced, heterogeneous
- Small size in each group, large numbers in total

Linear Mixed Models(LMMS) and REstricted Maximum Likelihood Methods(REML)

▷ Background
Motivation
▷ LMMs & REML
Problems
Challenges
Methods
Results
Discussion

LMMs: conceptually simple
REML: less biased estimates, widely used

Linear Mixed Models(LMMS) and REstricted Maximum Likelihood Methods(REML)

▷ Background
Motivation
▷ LMMs & REML
Problems
Challenges
Methods
Results
Discussion

LMMs: conceptually simple
REML: less biased estimates, widely used

$$y \approx X\tau$$

Observations $\approx$ fixed effects

Linear Mixed Models(LMMS) and REstricted Maximum Likelihood Methods(REML)

▷ Background
Motivation
▷ LMMs & REML
Problems
Challenges
Methods
Results
Discussion

LMMs: conceptually simple
REML: less biased estimates, widely used

$$y \approx X\tau + Zu$$

Observations $\approx$ fixed effects $+$random effects

Linear Mixed Models(LMMS) and REstricted Maximum Likelihood Methods(REML)

▷ Background
Motivation
▷ LMMs & REML
Problems
Challenges
Methods
Results
Discussion

LMMs: conceptually simple
REML: less biased estimates, widely used

$$y = X\tau + Zu + \epsilon$$

Observations = fixed effects +random effects +noise

Linear Mixed Models(LMMS) and REstricted Maximum Likelihood Methods(REML)

▷ Background
Motivation
▷ LMMs & REML
Problems
Challenges
Methods
Results
Discussion

LMMs: conceptually simple
REML: less biased estimates, widely used

$$y = X\tau + Zu + \epsilon$$

Observations = fixed effects +random effects +noise

$$E(u) = 0, E(\epsilon) = 0, u \sim N(0, \sigma^2 G), \epsilon \sim N(0, \sigma^2 R)$$

$$\text{var} \begin{pmatrix} u \\ \epsilon \end{pmatrix} = \sigma^2 \begin{pmatrix} G & \\ & R \end{pmatrix}$$

LMMS: Forward and Inverse Problems

▷ Background
Motivation
LMMs & REML
▷ Problems
Challenges
Methods
Results
Discussion

Given y, G and R find τ and u, via $y = X\tau + Zu + \epsilon$

LMMS: Forward and Inverse Problems

▷ Background
Motivation
LMMs & REML
▷ Problems
Challenges
Methods
Results
Discussion

Given y, G and R find τ and u, via $y = X\tau + Zu + \epsilon$
Ordinary Least Square or REML,

$$\underbrace{\begin{pmatrix} X^T R^{-1} X & X^T R^{-1} Z \\ Z^T R^{-1} X & Z^T R^{-1} Z + G^{-1} \end{pmatrix}}_{C} \begin{pmatrix} \hat{\tau} \\ \tilde{u} \end{pmatrix} = \begin{pmatrix} X^T R^{-1} y \\ Z^T R^{-1} y \end{pmatrix}$$

LMMS: Forward and Inverse Problems

▷ Background
Motivation
LMMs & REML
▷ Problems
Challenges
Methods
Results
Discussion

Given y, G and R find τ and u, via $y = X\tau + Zu + \epsilon$
Ordinary Least Square or REML,

$$\underbrace{\begin{pmatrix} X^T R^{-1} X & X^T R^{-1} Z \\ Z^T R^{-1} X & Z^T R^{-1} Z + G^{-1} \end{pmatrix}}_{C} \begin{pmatrix} \hat{\tau} \\ \tilde{u} \end{pmatrix} = \begin{pmatrix} X^T R^{-1} y \\ Z^T R^{-1} y \end{pmatrix}$$

The right question is how confident the estimates are.

$$\mathrm{var} \begin{pmatrix} \hat{\tau} - \tau \\ \tilde{u} - u \end{pmatrix} = \sigma^2 \begin{pmatrix} X^T R^{-1} X & X^T R^{-1} Z \\ Z^T R^{-1} X & Z^T R^{-1} Z + G^{-1} \end{pmatrix}^{-1}$$

LMMS: Forward and Inverse Problems

▷ Background
Motivation
LMMs & REML
▷ Problems
Challenges
Methods
Results
Discussion

Given y, G and R find τ and u, via $y = X\tau + Zu + \epsilon$
Ordinary Least Square or REML,

$$\underbrace{\begin{pmatrix} X^T R^{-1} X & X^T R^{-1} Z \\ Z^T R^{-1} X & Z^T R^{-1} Z + G^{-1} \end{pmatrix}}_{C} \begin{pmatrix} \hat{\tau} \\ \tilde{u} \end{pmatrix} = \begin{pmatrix} X^T R^{-1} y \\ Z^T R^{-1} y \end{pmatrix}$$

The right question is how confident the estimates are.

$$\mathrm{var} \begin{pmatrix} \hat{\tau} - \tau \\ \tilde{u} - u \end{pmatrix} = \sigma^2 \begin{pmatrix} X^T R^{-1} X & X^T R^{-1} Z \\ Z^T R^{-1} X & Z^T R^{-1} Z + G^{-1} \end{pmatrix}^{-1}$$

Only diagonal elements of C^{-1} are of great importance.

- Even C is sparse, C^{-1} is dense, Duff, Erisman, Gear & Reid(88)
- Compute $\mathrm{diag}(C^{-1})$, Erisman & Tinney(75), Cambell& Davis(95), Amestoy & Duff et al (2012)

LMMS: Forward and Inverse Problems

▷ Background
Motivation
LMMs & REML
▷ Problems
Challenges
Methods
Results
Discussion

Given y, G and R find τ and u, via $y = X\tau + Zu + \epsilon$
Ordinary Least Square or REML,

$$\underbrace{\begin{pmatrix} X^T R^{-1} X & X^T R^{-1} Z \\ Z^T R^{-1} X & Z^T R^{-1} Z + G^{-1} \end{pmatrix}}_{C} \begin{pmatrix} \hat{\tau} \\ \tilde{u} \end{pmatrix} = \begin{pmatrix} X^T R^{-1} y \\ Z^T R^{-1} y \end{pmatrix}$$

The right question is how confident the estimates are.

$$\mathrm{var}\begin{pmatrix} \hat{\tau} - \tau \\ \tilde{u} - u \end{pmatrix} = \sigma^2 \begin{pmatrix} X^T R(\phi)^{-1} X & X^T R(\phi)^{-1} Z \\ Z^T R(\phi)^{-1} X & Z^T R(\phi)^{-1} Z + G(\gamma)^{-1} \end{pmatrix}^{-1}$$

From real problems, the co-variance matrices R and G are unknown or $R(\phi)$ and $G(\gamma)$.

Find $R(\phi)$ and $G(\gamma)$ first.

Problem

▷ Background
Motivation
LMMs & REML
▷ Problems
Challenges
Methods
Results
Discussion

(Restricted) Maximum Likelihood Principle:

$$\ell_R(\theta) = \text{const} - \frac{1}{2}\{(n-\nu)\log\sigma^2 + \log\det(H) + \log\det(X^T H^{-1} X) + \frac{y^T P y}{\sigma^2}\}, \quad (1)$$

where, $V = \sigma^2\left(R(\phi) + ZG(\gamma)Z^T\right) := \sigma^2 H$

$$P = H^{-1} - H^{-1}X(X^T H^{-1} X)^{-1} X H^{-1}$$

$\theta = (\sigma^2; \phi; \gamma)$

Problem

▷ Background
Motivation
LMMs & REML
▷ Problems
Challenges
Methods
Results
Discussion

(Restricted) Maximum Likelihood Principle:

$$\ell_R(\theta) = \text{const} - \frac{1}{2}\{(n-\nu)\log\sigma^2 + \log\det(H) + \log\det(X^T H^{-1} X) + \frac{y^T P y}{\sigma^2}\}, \quad (1)$$

where, $V = \sigma^2\left(R(\phi) + ZG(\gamma)Z^T\right) := \sigma^2 H$

$$P = H^{-1} - H^{-1}X(X^T H^{-1} X)^{-1} X H^{-1}$$

$\theta = (\sigma^2; \phi; \gamma)$ Our task:

$$\theta^* = \arg\max_{\theta} \ell_R(\theta)$$

Background
▷ Challenges
Challenges I
Challenge II
Methods
Results
Discussion

Challenges

Challenges I

Background
▷ Challenges
▷ Challenges I
Challenge II
Methods
Results
Discussion

1. The objective function is too complicated.
   logdet(complicated matrix products) terms.

Challenges I

Background
▷ Challenges
▷ Challenges I
Challenge II
Methods
Results
Discussion

1. The objective function is too complicated.
   logdet(complicated matrix products) terms.
2. For derivative methods(Newton Method), the observed information matrix is computationally prohibitive.

Challenges I

Background
▷ Challenges
▷ Challenges I
Challenge II
Methods
Results
Discussion

1. The objective function is too complicated.
   logdet(complicated matrix products) terms.
2. For derivative methods(Newton Method), the observed information matrix is computationally prohibitive.

Score function(1^{st} derivative of ℓ_R):

$$S(\theta_i) = \frac{\partial \ell_R}{\partial \theta_i} = -\frac{1}{2}\left\{\mathrm{tr}\left(P\dot{V}_i\right) - y^T P\dot{V}_j P y\right\}, \quad \dot{V}_i = \frac{\partial V}{\partial \theta_i}$$

Challenges I

Background
▷ Challenges
▷ Challenges I
Challenge II
Methods
Results
Discussion

1. The objective function is too complicated.
   logdet(complicated matrix products) terms.
2. For derivative methods(Newton Method), the observed information matrix is computationally prohibitive.

Score function(1^{st} derivative of ℓ_R):

$$S(\theta_i) = \frac{\partial \ell_R}{\partial \theta_i} = -\frac{1}{2}\left\{\mathrm{tr}\left(P\dot{V}_i\right) - y^T P\dot{V}_j P y\right\}, \quad \dot{V}_i = \frac{\partial V}{\partial \theta_i}$$

Newton Method to find $S(\theta) = 0$

1. Give a guess θ_0
2. For $k = 1$ till convergence
3. solve $J(\theta_k)\delta_k = -S(\theta_k)$
4. $\theta_{k+1} = \theta_k$
5. EndFor

Challenges I

Background
▷ Challenges
▷ Challenges I
Challenge II
Methods
Results
Discussion

1. The objective function is too complicated. logdet(complicated matrix products) terms.
2. For derivative methods(Newton Method), the observed information matrix is computationally prohibitive.

Score function(1^{st} derivative of ℓ_R):

$$S(\theta_i) = \frac{\partial \ell_R}{\partial \theta_i} = -\frac{1}{2}\left\{\mathrm{tr}\left(P\dot{V}_i\right) - y^T P\dot{V}_j P y\right\}, \quad \dot{V}_i = \frac{\partial V}{\partial \theta_i}$$

Observed information matrix(2^{nd} derivative of $-\ell_R$):

$$\mathcal{I}(\theta_i, \theta_j) = -J = -\frac{\partial \ell_R^2}{\partial \theta_i \partial \theta_j} =$$

$$\frac{1}{2}\left\{\mathrm{tr}(P\ddot{V}_{ij}) - \mathrm{tr}(P\dot{V}_i P\dot{V}_j) + 2y^T P\dot{V}_i P\dot{V}_j y - y^T P\ddot{V}_{ij} P y\right\}$$

$$\ddot{V}_{ij} = \frac{\partial V^2}{\partial \theta_i \partial \theta_j}$$

Challenge II

Background
▷ Challenges
Challenges I
▷ Challenge II
Methods
Results
Discussion

3 Quasi-Newton method: Fisher-Scoring Algorithm, simper but still not scalable

$$\underbrace{-J \approx \mathcal{I}_F(\theta_i, \theta_j) = \frac{1}{2}\mathrm{tr}(P\dot{V}_i P\dot{V}_j)}_{\text{Fisher information matrix}}$$

4 Derivative free methods: converges very slow or doesn't converge for some difficult problems.

☐ Misztal, Comparison of computing properties of derivative and derivative-free algorithms in variance component estimation by REML. Journal of Animal Breeding and Genetics, 111(1-6):346C355, 1994.

5 For iterative methods, there are plenty of research opportunities for longer time scale.

Background
Challenges
▷ Methods
AI
AIS
Matrix transforms
Sparse direct solver, when require ordering
Results
Discussion

Methods

Average information for $\dot{V}_{ij} = 0$

Background
Challenges
▷ Methods
▷ AI
AIS
Matrix transforms
Sparse direct solver, when require ordering
Results
Discussion

- Observed information matrix: $\mathcal{I}_O(\theta_i, \theta_j) = \frac{1}{2}\left\{\mathrm{tr}(P\ddot{V}_{ij}) - \mathrm{tr}(P\dot{V}_i P\dot{V}_j) + 2y^T P\dot{V}_i P\dot{V}_j Py - y^T P\ddot{V}_{ij} Py\right\}$
- The Fisher information matrix:$\mathcal{I}_F(\theta_i, \theta_j) = \frac{1}{2}\mathrm{tr}(P\dot{V}_i P\dot{V}_j)$

For the covariance matrices linearly depends on θ, i.e, $\ddot{V}_{ij} = 0$

Average information for $\dot{V}_{ij} = 0$

Background
Challenges
▷ Methods
▷ AI
AIS
Matrix transforms
Sparse direct solver, when require ordering
Results
Discussion

- Observed information matrix: $\mathcal{I}_O(\theta_i, \theta_j) = \frac{1}{2}\left\{\mathrm{tr}(P\ddot{V}_{ij}) - \mathrm{tr}(P\dot{V}_i P\dot{V}_j) + 2y^T P\dot{V}_i P\dot{V}_j Py - y^T P\ddot{V}_{ij} Py\right\}$
- The Fisher information matrix:$\mathcal{I}_F(\theta_i, \theta_j) = \frac{1}{2}\mathrm{tr}(P\dot{V}_i P\dot{V}_j)$
- Average information matrix:
  - Gilmour & Thompson & Cullis(95), Mayer(97)

$$\mathcal{I}_A = \frac{\mathcal{I}_O + \mathcal{I}_F}{2} = \frac{1}{2} y^T P\dot{V}_i P\dot{V}_j Py$$

For the covariance matrices linearly depends on θ, i.e, $\ddot{V}_{ij} = 0$

Average Information Splitting for $\ddot{V}_{ij} \neq 0$

Background
Challenges
▷ Methods
AI
▷ AIS
Matrix transforms
Sparse direct solver, when require ordering
Results
Discussion

Theorem 1. *(Zhu et al 2016) Let $\mathcal{I}_O$ and $\mathcal{I}$ be the observed information and the Fisher information for the restricted log-likelihood of the LMM respectively, then the average of the observed and Fisher information can be split as*

$$\frac{\mathcal{I}_O + \mathcal{I}}{2} = \mathcal{I}_A + \mathcal{I}_Z$$

such that the expectation of $\mathcal{I}_A$ is the Fisher information and $E(\mathcal{I}_Z) = 0$

$$\mathcal{I}_A = \frac{\mathcal{I}_O + \mathcal{I}_F}{2} = \frac{1}{2} y^T P \dot{V}_i P \dot{V}_j P y \qquad \text{simple, essential} \qquad (2)$$

$$\mathcal{I}_Z = \frac{1}{4}\{\mathrm{tr}(P\ddot{V}_{ij}) - y^T P^T \ddot{V}_{ij} P y\} \qquad \text{complicated, negligible} \qquad (3)$$

$\mathrm{tr}(P\dot{V}_i P\dot{V}_j)$ to a quadratic form (5 SpMVs + 1 dot)!.

Matrix transforms

Background
Challenges
▷ Methods
AI
AIS
Matrix
▷ transforms
Sparse direct solver, when require ordering
Results
Discussion

How to comput Py, where
$P = H^{-1} - H^{-1}X(X^TH^{-1}X)^{-1}XH^{-1}$?
$\left(R(\phi) + ZG(\gamma)Z^T\right) = H$

Theorem 2. *$Py = R^{-1}e$, where $e = y - X\hat{\tau} - Z\tilde{u}$, $\hat{tau}$ and $\tilde{u}$ are the solution to the mixed model equation*

$$\begin{pmatrix} X^TR^{-1}X & X^TR^{-1}Z \\ Z^TR^{-1}X & Z^TR^{-1}Z + G^{-1} \end{pmatrix} \begin{pmatrix} \hat{\tau} \\ \tilde{u} \end{pmatrix} = \begin{pmatrix} X^TR^{-1}y \\ Z^TR^{-1}y \end{pmatrix} \quad (4)$$

- H is $n \times n$, n the number of observations
- C is $(p+b) \times (p+b)$, the number of fixed and random effects.
- n can be $\gg p + b$ for many cases
- R is the co-variance structure for the noise, usually diagonal or block diagonal, easy to invert.

Matrix transforms

Background
Challenges
▷ Methods
AI
AIS
Matrix
▷ transforms
Sparse direct solver, when require
ordering
Results
Discussion

Where linear system with multiple right hands come from?

Let $Y_i = \dot{V}_i P y$, $PY_i = P\dot{V}_i P y$. Let $Y = (Y_1, Y_2, \cdots Y_m)$.

$PY = R^{-1}E$, where $E = y - X\hat{T} - Z\tilde{U}$, $\hat{T}$ and $\tilde{U}$ are the solution to the mixed model equation

$$\begin{pmatrix} X^T R^{-1} X & X^T R^{-1} Z \\ Z^T R^{-1} X & Z^T R^{-1} Z + G^{-1} \end{pmatrix} \begin{pmatrix} \hat{T} \\ \tilde{U} \end{pmatrix} = \begin{pmatrix} X^T R^{-1} Y \\ Z^T R^{-1} Y \end{pmatrix} \quad (4)$$

Matrix transforms

Background
Challenges
▷ Methods
AI
AIS
Matrix
▷ transforms
Sparse direct solver, when require
ordering
Results
Discussion

Where linear system with multiple right hands come from?

Let $Y_i = \dot{V}_i P y$, $PY_i = P\dot{V}_i P y$. Let $Y = (Y_1, Y_2, \cdots Y_m)$.

$PY = R^{-1}E$, where $E = y - X\hat{T} - Z\tilde{U}$, $\hat{T}$ and $\tilde{U}$ are the solution to the mixed model equation

$$\begin{pmatrix} X^T R^{-1} X & X^T R^{-1} Z \\ Z^T R^{-1} X & Z^T R^{-1} Z + G^{-1} \end{pmatrix} \begin{pmatrix} \hat{T} \\ \tilde{U} \end{pmatrix} = \begin{pmatrix} X^T R^{-1} Y \\ Z^T R^{-1} Y \end{pmatrix} \quad (4)$$

Compute $\mathcal{I}_A = y^T P\dot{V}_i P\dot{V}_j P y$

1. $\xi = Py$
2. $Y_i = P\dot{V}_i\xi$, solve linear system with only one right hand side.
3. $\eta_j = PY_i$, solve linear systems with multiple right hand sides.
4. $\mathcal{I}(\theta_i, \theta_j) = Y_i^T \eta_j$

Sparse direct solver, when require

Background
Challenges
▷ Methods
AI
AIS
Matrix transforms
Sparse direct solver, when
▷ require
ordering
Results
Discussion

I. Compute $\mathcal{I}_A = y^T P\dot{V}_i P\dot{V}_j Py$

1. $\xi = Py$ solve linear system with only one RHS.
2. $Y_i = \dot{V}_i \xi$,
3. $\eta_j = PY_i$, solve linear systems with multiple RHSs.
4. $\mathcal{I}(\theta_i, \theta_j) = Y_i^T \eta_j$

II In Newton-iterative solvers

1. Give a guess θ_0
2. For $k = 1$ till convergence
3. solve $\mathcal{I}_A(\theta_k)\delta_k = S(\theta_k)$
4. $\theta_{k+1} = \theta_k$
5. EndFor

III. When have $R(\phi)$, $G(\gamma)$, find the diagonal entries of C^{-1}

IV. Evaluate the likelihood function (equivalent to $\log\det(C) + \log(R) + \log(G) + \cdots$)

Sparse matrix techniques

Background
Challenges
▷ Methods
AI
AIS
Matrix transforms
Sparse direct solver, when require
▷ ordering
Results
Discussion

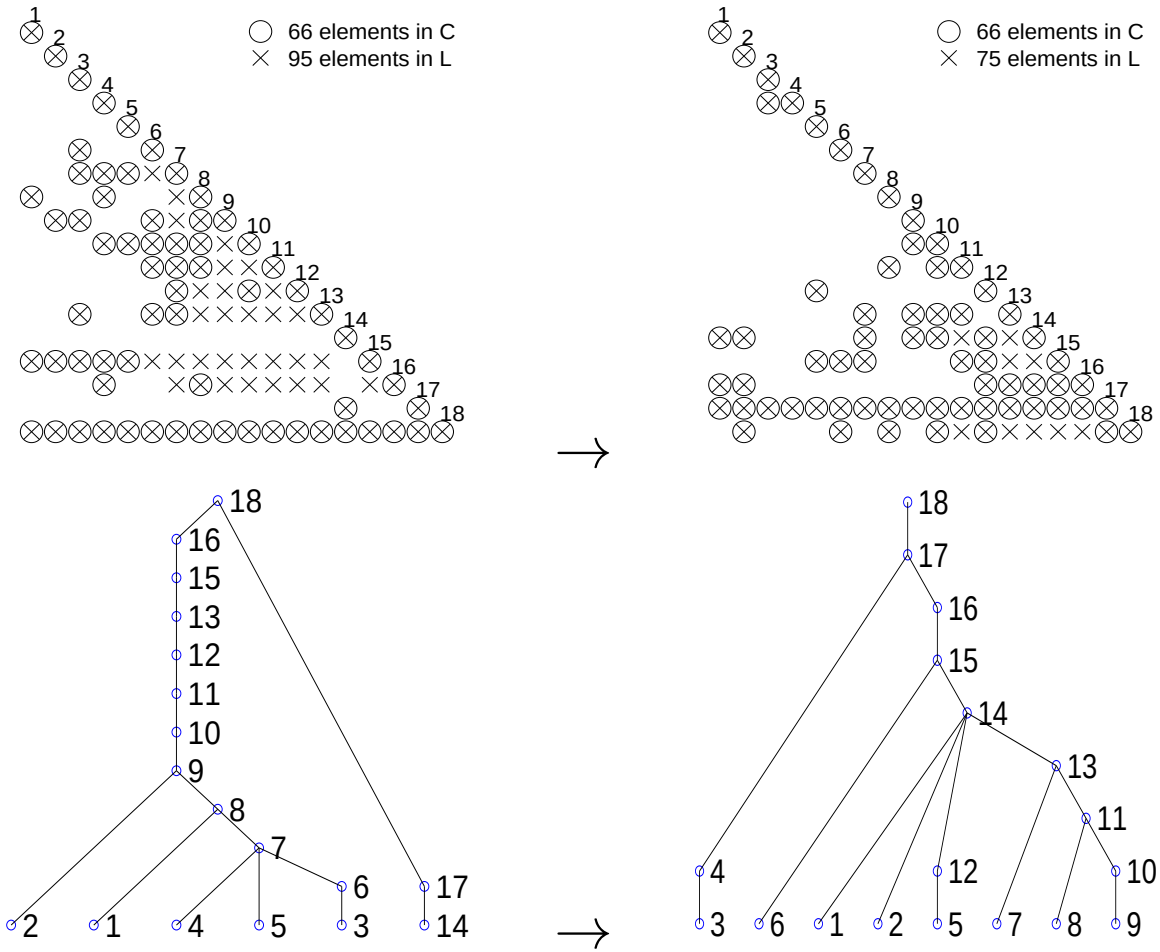
1. Symbolic analysis for predicting the structure of L, static memory management rather than dynamic memory allocation.
2. Fill-in reducing ordering to reduce computations and increase parallelism. Vital for this application, since $C = LDL^T$ are repeatedly used.
   - Duff(81), Duff & Wiberg , ACMTMS(88), Duff & Meurant, BIT(89), Amestoy & Davis & Duff, AMD, SIAMX(96) Duff & Kaya & Ucar, ACMTMS (2011), several talks here
3. Multi-frontal solvers , the LDL^T version is used.
   - Davis, Algorithm 849, Duff & Reid (83)
4. Inverse multi-frontal methods for selected inversion algorithm was implemented by us.
   - Davis (95), Lin Lin et al, Selinv (2012), Amestoy, Duff et al
   - unpublished `selinv03`

Background
Challenges
Methods
▷ Results
Demo
Benchmark problems
More examples
Discussion

Results

Demo

Background
Challenges
Methods
▷ Results
▷ Demo
Benchmark problems
More examples
Discussion



Benchmark problems

Background
Challenges
Methods
▷ Results
Demo
Benchmark problems
More examples
Discussion

DataSet	y	c	v	y.c	y.v	v.c	units	v/y	y/v	c.v
prob_1	12	22	130	132	673	2518	6667	56.1	5.2	10
prob_2	15	25	160	180	888	3527	9595	59.2	5.6	10
prob_3	22	25	188	264	1177	4215	12718	53.5	6.3	12
prob_4	25	25	262	300	1612	5907	17420	64.5	6.2	12
prob_5	25	25	390	300	2345	8625	25334	93.8	6.0	15
prob_6	25	35	390	425	2345	12249	35887	93.8	6.0	15
prob_7	30	35	470	510	3013	15087	46113	100.4	6.4	20
prob_8	30	35	620	510	3835	19737	58685	127.8	6.2	20
prob_9	35	40	720	700	4522	26432	81396	129.2	6.3	20
prob_10	40	50	820	1000	5262	37701	118403	131.6	6.4	20

Table 1: Plant breeding data

Benchmark problems

Background
Challenges
Methods
▷ Results
Demo
Benchmark
▷ problems
More examples
Discussion

Prob	No. effects	sparsity of C			Sparsity of L			FLOPs count	
		nnz	n_z	per 1000	nnz	n_z	per 1000	LDL^T	Selinv
prob_01	3488	56946	16.3	9.4	112618	32.3	18.5	8943842	17778554
prob_02	4796	80946	16.9	7.0	172023	35.9	15.0	17175555	34183883
prob_03	5892	105059	17.8	6.1	273315	46.4	15.7	40768817	81270211
prob_04	8132	144240	17.7	4.4	377761	46.5	11.4	60714709	121059789
prob_05	11711	209235	17.9	3.1	507711	43.4	7.4	75897428	151298856
prob_06	15470	291318	18.8	2.4	718701	46.5	6.0	149074099	297444967
prob_07	19146	370799	19.4	2.0	1020414	53.3	5.6	270835518	540669768
prob_08	24768	473891	19.1	1.5	1196903	48.3	3.9	290699965	580227795
prob_09	32450	648237	20.0	1.2	1779662	54.8	3.4	600925570	1200103928
prob_10	44874	932054	20.8	0.9	2817463	62.8	2.8	1391099157	2779425725

$$\underbrace{\begin{pmatrix} X^TR^{-1}X & X^TR^{-1}Z \\ Z^TR^{-1}X & Z^TR^{-1}Z + G^{-1} \end{pmatrix}}_{C} \begin{pmatrix} \hat{\tau} \\ \tilde{u} \end{pmatrix} = \begin{pmatrix} X^TR^{-1}y \\ Z^TR^{-1}y \end{pmatrix} \quad (5)$$

Benchmark problems

Background
Challenges
Methods
▷ Results
Demo
Benchmark
▷ problems
More examples
Discussion

Table 1: Speed up over VSNi's existing software

prob	AMD	LDL^T	Selinv	Factor+Inv	ALL
ex_2	-	25.98	-	10.51	6.46
ex_3	15.41	-	-	-	4.95
ex_4	14.78	-	5.34	3.76	7.55
ex_5	12.32	-	5.77	3.66	6.65
ex_6	15.15	8.91	2.21	3.86	7.16
ex_7	17.25	12.67	2.46	4.71	8.17
ex_8	19.01	11.01	2.31	4.41	9.22
ex_9	26.29	14.92	2.29	4.73	10.64
ex_10	36.78	5.93	2.46	3.85	9.31

More examples

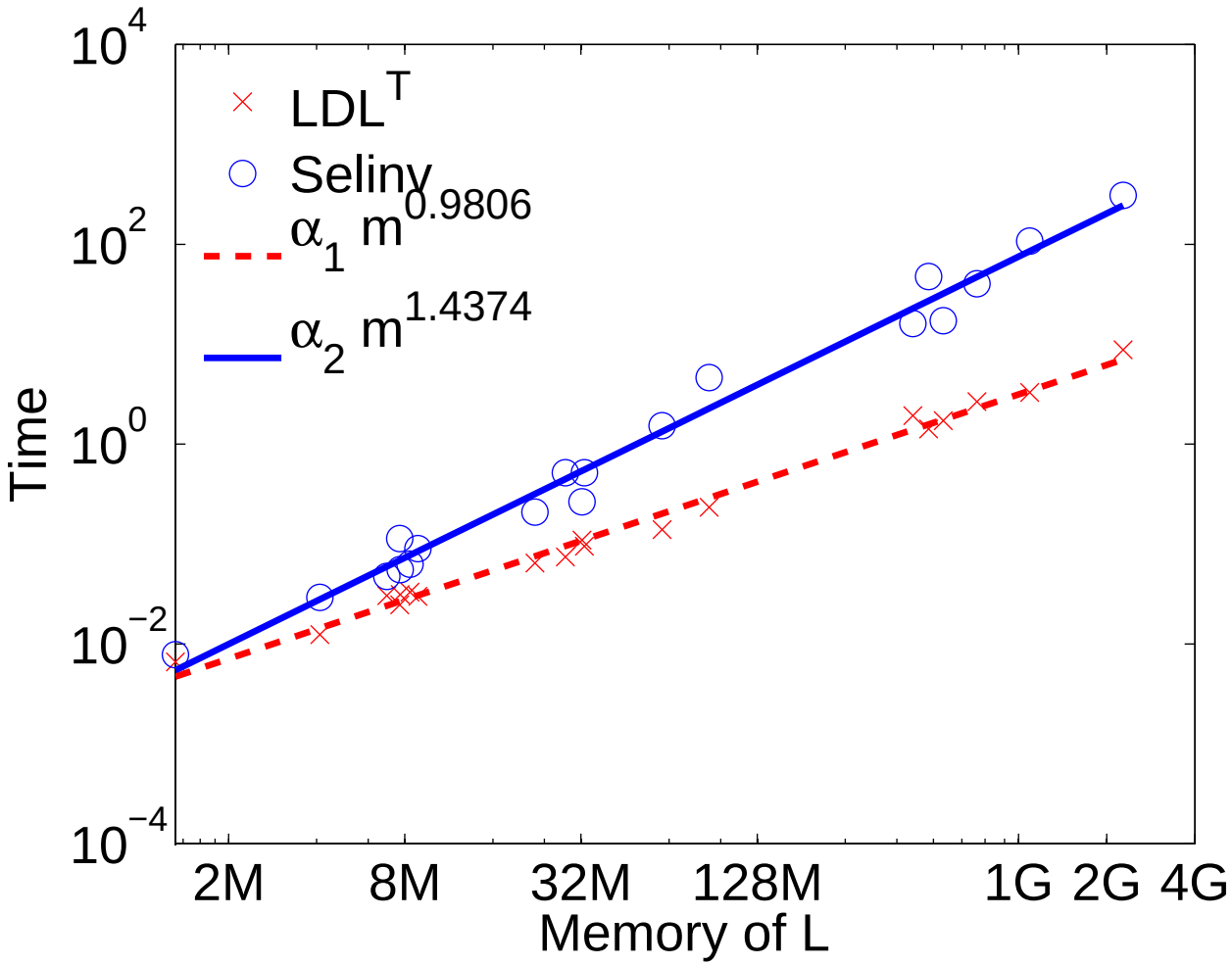
Background
Challenges
Methods
▷ Results
Demo
Benchmark problems
▷ More examples
Discussion

Prob	AMD	LDL^T	selinv3o	LDU	Selinv
HB/bcsstk14	0.0009	0.0066	0.0078	0.0194	0.0372
HB/bcsstk28	0.0017	0.0124	0.0293	0.0712	0.1152
Boeing/bcsstk38	0.0045	0.0300	0.0990	0.1487	0.3044
HB/bcsstk18	0.0074	0.0246	0.1136	0.1482	0.4091
TKK/cbuckle	0.0088	0.0743	0.5180	0.4660	1.7940
Pothen/bodyy4	0.0059	0.0305	0.0474	0.1097	0.1960
Pothen/bodyy5	0.0059	0.0313	0.0556	0.1251	0.2158
Pothen/bodyy6	0.0061	0.0330	0.0629	0.1364	0.2444
Simon/raefsky4	0.0147	0.2338	4.6353	1.3349	11.5883
Boeing/bcsstk36	0.0108	0.0954	0.5170	0.5379	1.8475
Boeing/crystm03	0.0137	0.1399	1.5345	0.7527	3.3403
GHS_psdef/wathen120	0.0095	0.0645	0.2095	0.4088	1.1996
Schmid/thermal1	0.0501	0.1096	0.2645	0.5538	1.0480
DNVS/shipsec1	0.2250	1.4231	47.5587	8.8262	134.2235
Boeing/pwtk	0.1640	2.6615	40.3312	13.0838	131.9626
Wissgott/parabolic_fem	0.4778	1.9287	16.1115	7.2807	48.4882
McRae/ecology2	0.4012	1.7227	17.2147	9.9427	75.0454
GHS_psdef/bmwcra_1	0.1618	3.2910	107.9326	111.6025	428.1022
AMD/G3_circuit	1.5675	8.8011	309.9106	out of	memory

- Duff & Grimes & Lewis(92) Harwell-Boeing Sparse Matrix Collection,
- Davis , SF collection

More examples

Background
Challenges
Methods
▷ Results
Demo
Benchmark problems
▷ More examples
Discussion



Background
Challenges
Methods
Results
▷ Discussion
Discussion and Q&A

Discussion

Discussion and Q&A

Background
Challenges
Methods
Results
▷ Discussion
Discussion and
▷ Q&A

- The average information splitting techniques and dedicated matrix transform can significantly reduce computations, while the sparse matrices techniques real serve as kernels
- A prefect application which repeatedly employs the sparse director solvers (RSHs), inverse multi-frontal inversion algorithm (Selinv), reordering.
- As a whole, the sparse direct methods are preferred at moment.
- With increasing importance.

Discussion and Q&A

Background
Challenges
Methods
Results
▷ Discussion
Discussion and
▷ Q&A

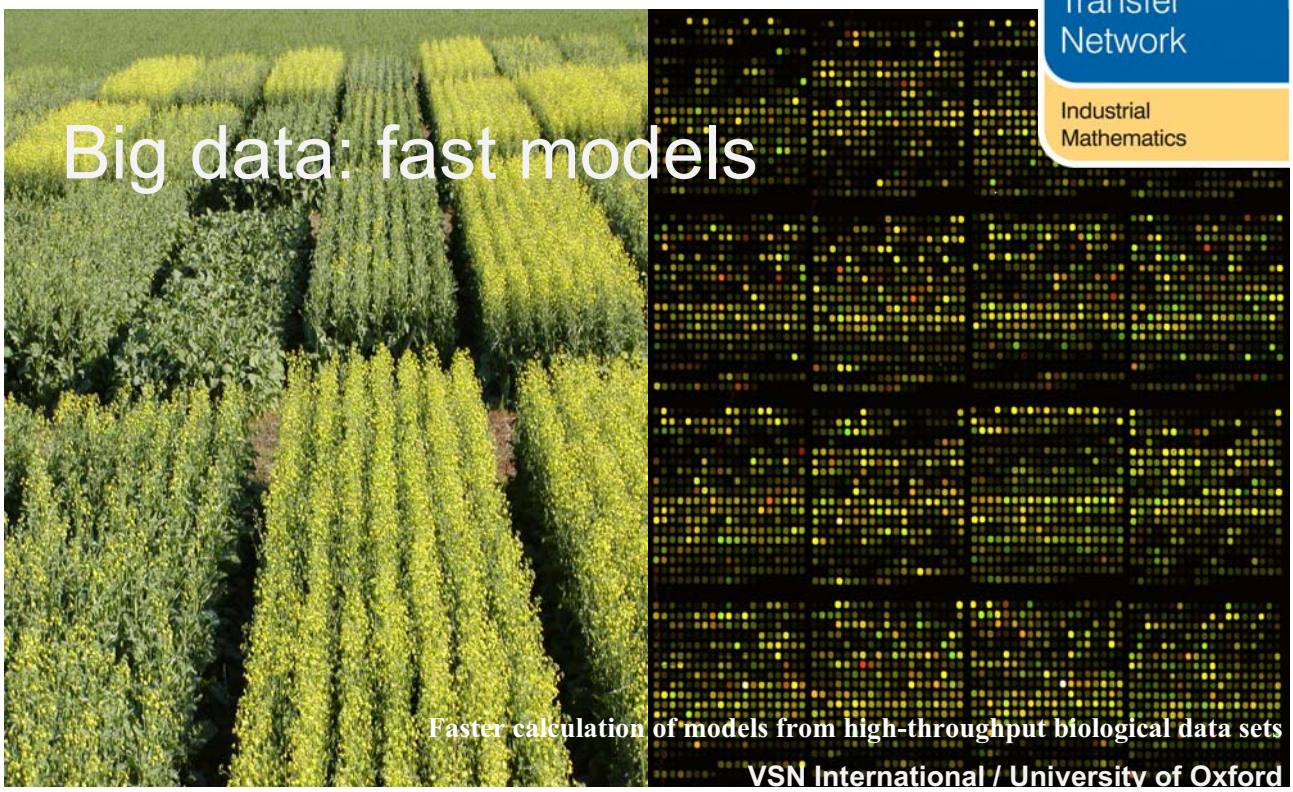
- The average information splitting techniques and dedicated matrix transform can significantly reduce computations, while the sparse matrices techniques real serve as kernels
- A prefect application which repeatedly employs the sparse director solvers (RSHs), inverse multi-frontal inversion algorithm (Selinv), reordering.
- As a whole, the sparse direct methods are preferred at moment.
- With increasing importance.

Thanks to Natural Science Foundation of China, Natural Science Foundation of Jiangsu Province.

Technology Strategy Board
Driving Innovation

Knowledge Transfer Network
Industrial Mathematics

Big data: fast models



Faster calculation of models from high-throughput biological data sets

VSN International / University of Oxford

The need

High-throughput technologies in the biological sciences mean that the sizes of data sets – and the corresponding statistical models – have suddenly increased by several orders of magnitude. For instance, a crop trial can now be analysed by a microarray.

Software packages that fit models to these types of data must find new ways to operate to remain competitive. Possibilities for exploiting new algorithms come from advances in computational mathematics and computer hardware (multi-threading and GPU).

VSN International is committed to improving its products to serve the biological sciences. We are investing in this research to ensure that we have state-of-the-art algorithms that will help us to increase our market share.

The outcomes

This short KTP project has enabled VSNi to obtain a review of the relevant mathematical methods, to evaluate some of these methods and to obtain a new algorithm for matrix inversion that is specifically tailored to our problems. This new capability can now be evaluated in the context of our current software and used to develop faster, more efficient methods.

Close collaboration of the VSNi Technical Team with the Associate has helped to improve understanding of the methods within the company, which is crucial to extending them to new scenarios. The Associate has also enabled better access to the mathematical literature and utility software packages. This has broadened the knowledge of the Technical Team and provides a good basis for moving forward.

The Associate has been able to experience working in a commercial environment, with practical constraints on investment and competing demands on time. He has also been able to experience working as part of a multi-project team and learn about the importance of good communication and best practice in software development.

This project provides an excellent example of the transfer of academic mathematical knowledge into a practical application within a commercial environment.

"Working with Shengxin has enabled VSNi to gain more knowledge of computational and mathematical methods of relevance to our products. The project has been of great value in bringing us up to speed in a highly specialist area ."

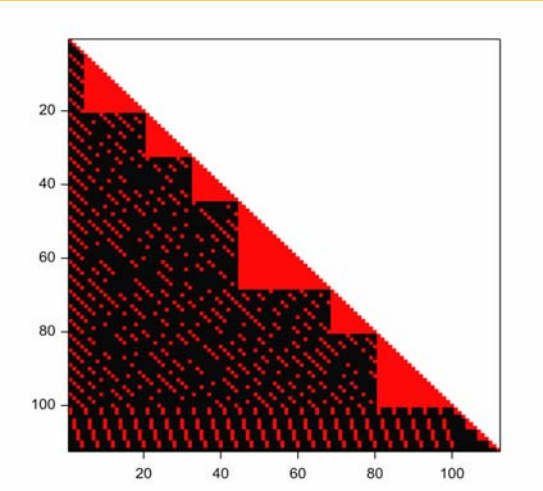
Sue Welham, Statistical Software Developer, VSN International

Technical summary

VSNi produce statistical software and specialise in algorithms for analysing linear mixed models. The method for fitting these models requires inversion of symmetric matrices which may be very large (many thousands of rows). In many standard biological problems, these matrices are sparse, with a large proportion of their entries being zero. Computational methods for sparse matrices take advantage of the fact that there is no need to include the zero entries in making calculations, thus greatly reducing the both the number of operations performed and the computing time required.

Modern computer architectures provide different methods of processing calculations in parallel (multi-threading, GPU). There are methods for inversion of dense matrices (i.e. matrices with only few zero entries) which take advantage of this architecture, but fewer algorithms for sparse matrices. In addition, the inverse of a sparse matrix is often relatively dense. VSNi software makes further savings by calculating only certain elements of the inverse matrix, but this modification requires specialist algorithms.

The aim of this shorter KTP project has been to review the literature, to assess the availability and viability of standard algorithms for parallel inversion of sparse symmetric matrices, and to develop extensions of these algorithms tailored to the specific needs of VSNi. One a suite of test examples the software dedeveloped in this project (from permuting to inversion) ran 10 times faster than VSNi's existing software.



Sparsity pattern of a small matrix typical of those VSNi needs to deal with. Only the lower triangle is shown, with zeros as black and non-zeros red.

"It is a great pleasure and opportunity to work with Dr. Sue Welham, Mr. Simon Harding and Prof. Robin Thompson on VSNi's real world problems. Their penetrating questions and rich software development experience led me to learn a lot at extreme speed in the last 6 months."

Shengxin Zhu, University of Oxford

"The difficulties in problems from industry are not always what is initially perceived, but with broad enough expertise, the academic contribution can still be very valuable as was the case for this project."

Andy Wathen, University of Oxford

This project was part of the programme of industrial mathematics shorter KTPs managed by the Knowledge Transfer Network (KTN) for Industrial Mathematics. The KTN works to exploit mathematics as an engine for innovation. It is supported by the Technology Strategy Board, in its role as the UK's national innovation agency, and the Engineering and Physical Sciences Research Council, in its role as the main UK government agency for funding research and training in engineering and the physical sciences.

EPSRC
Engineering and Physical Sciences Research Council

Project Details

Partners
VSN International Ltd
University of Oxford

Project investment
£10,000

Intern
Shengxin Zhu

For details on the technology:
Roger Payne
Chief Science & Technology Officer
VSN International
Hemel Hempstead, HP1 1ES
roger.payne@vsni.co.uk
+44 01442 450230

For information on internships and other collaborations:
Lorcán Mac Manus
Industrial Mathematics KTN
lbmm@industrialmaths.net
+44 (0) 1483 565252