Sparse Matrices
and Large Data Issues

Workshop ENAR

March 15, 2009

Reinhard Furrer, CSM
Outline

What are sparse matrices?

How to work with sparse matrices?

Sparse positive definite matrices in statistics.

Sparse matrices and fields.
Sparse Matrices

What is “sparse” or a sparse matrix?

According to Wiktionary/Wikipedia:

Sparse: (Adjective)
1. Having widely spaced intervals
2. Not dense; meager

Sparse matrix:
a matrix populated primarily with zeros.

Sparse Matrices

R> n <- 15
R> A <- array( runif(n^2), c(n,n)) + diag(n)
R> A[A < 0.75] <- 0

```
row
2 4 6 8 10 12 14
14 12 10 8 6 4 2
```
Sparse Matrices

R> n <- 15
R> A <- array( runif(n^2), c(n,n)) + diag(n)
R> A[A < 0.75] <- 0
R> AtA <- t(A) %*% A

Why should we use sparse matrices?
1. Savings in storage
2. Savings in computing time
Sparse Matrices

Why should we use sparse matrices?

1. Savings in storage
   nonzeros vs total elements

2. Savings in computing time
   0.066s vs 0.003 for 1,000 × 1,000 matrix multiplication

To exploit the savings need to exploit the sparsity.

We need a clever storage format and fast algorithms.
Storage Formats

Let $\mathbf{A} = (a_{ij}) \in \mathbb{R}^{n \times m}$ and $z$ the number of its nonzero elements.

1. Naive/“traditional”/classic format:
   one vector of length $n \times m$ and a dimension attribute.

2. Triplet format:
   three vectors of length $z$, $(i, j, a_{ij})$ and a dimension attribute.

3. Compressed sparse row (CSR) format:
   eliminate redundant row indices.

4. and about 10 more . . .
Storage Formats, Example

\[ A = \begin{pmatrix}
1 & 0.1 & 0 & 0.2 & 0.3 \\
0.4 & 2 & 0 & 0.5 & 0 \\
0 & 0 & 3 & 0 & 0.6 \\
0.7 & 0.8 & 0 & 4 & 0 \\
0.9 & 0 & 0.0 & 0 & 5 \\
\end{pmatrix} \]

Naive/traditional/classic:
1, .4, 0, .7, .9, .1, 2, 0, .8, 0, 0, 0, 3, 0, .0, .2, .5, 0, 4, 0, .3, 0, .6, 0, 5
Storage Formats, Example

\[ A = \begin{pmatrix}
  1 & 0.1 & 0 & 0.2 & 0.3 \\
  0.4 & 2 & 0 & 0.5 & 0 \\
  0 & 0 & 3 & 0 & 0.6 \\
  0.7 & 0.8 & 0 & 4 & 0 \\
  0.9 & 0 & 0.0 & 0 & 5
\end{pmatrix} \]

Triplet:

\[ i : \quad 1 \quad 1 \quad 1 \quad 1 \quad 2 \quad 2 \quad 2 \quad 3 \quad 3 \quad 4 \quad 4 \quad 4 \quad 5 \quad 5 \quad 5 \quad 5 \]

\[ j : \quad 1 \quad 2 \quad 4 \quad 5 \quad 1 \quad 2 \quad 4 \quad 2 \quad 3 \quad 1 \quad 2 \quad 4 \quad 1 \quad 3 \quad 5 \]

\[ a_{ij} : \quad 1 \quad .1 \quad .2 \quad .3 \quad .4 \quad 2 \quad .5 \quad 3 \quad .6 \quad .7 \quad .8 \quad 4 \quad .9 \quad .0 \quad 5 \]
Storage Formats, Example

\[
A = \begin{pmatrix}
1 & 0.1 & 0 & 0.2 & 0.3 \\
0.4 & 2 & 0 & 0.5 & 0 \\
0 & 0 & 3 & 0 & 0.6 \\
0.7 & 0.8 & 0 & 4 & 0 \\
0.9 & 0 & 0.0 & 0 & 5
\end{pmatrix}
\]

\[
\begin{array}{c}
i : \\
j : \\
a_{ij} : \\
\end{array}
\begin{array}{c}
1 \\
1 \ 2 \ 4 \ 5 \\
1 \ .1 \ .2 \ .3 \ .4 \ .2 \ .5 \\
\end{array}
\begin{array}{ccccc}
2 & 3 & 4 & 5 \\
1 \ 2 \ 4 & 3 & 1 \ 2 \ 4 & 1 \ 3 \ 5 \\
.2 \ .3 \ .4 \ .5 & .6 & .7 \ .8 \ .4 \ .9 \ .0 \ 5
\end{array}
Storage Formats, Example

\[ A = \begin{pmatrix}
1 & 0.1 & 0 & 0.2 & 0.3 \\
0.4 & 2 & 0 & 0.5 & 0 \\
0 & 0 & 3 & 0 & 0.6 \\
0.7 & 0.8 & 0 & 4 & 0 \\
0.9 & 0 & 0.0 & 0 & 5
\end{pmatrix} \]

CSR:
\[ ptr : \] 1 5 8 13 16
\[ j : \] 1 2 4 5 1 2 4 2 3 1 2 4 1 3 5
\[ a_{ij} : \] 1 .1 .2 .3 .4 2 .5 3 .6 .7 .8 4 .9 .0 5
Compressed Sparse Row Format

1. the nonzero values row by row

2. the (ordered) column indices of nonzero values

3. the position in the previous two vectors corresponding to new rows, given as pointers

4. the column dimension of the matrix.

CSR:

<table>
<thead>
<tr>
<th>ptr</th>
<th>1</th>
<th>2</th>
<th>4</th>
<th>5</th>
<th>8</th>
<th>10</th>
<th>13</th>
<th>16</th>
</tr>
</thead>
<tbody>
<tr>
<td>j</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>5</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>a_{ij}</td>
<td>1</td>
<td>.1</td>
<td>.2</td>
<td>.3</td>
<td>.4</td>
<td>2</td>
<td>.5</td>
<td>3</td>
</tr>
</tbody>
</table>
(Dis)Advantages

1. Naive format:
   No savings in storage and computation (for sparse matrices)
   Status quo

2. Triplet format:
   Savings in storage and computation for sparse matrices
   Loss in storage and computation for full matrices
   Intuitive

3. Compressed sparse row (CSR) format:
   Apart from intuitive, same as triplet
   Faster element access
   Many available algorithms
   Arbitrary choice for “row” vs “column” format (CSC)
Implications

With a new storage format new “algorithms” are required . . .

Is it worthwhile???
Timing

Setup:

R> timing <- function(expr)
   +   as.vector( system.time( for (i in 1:N) expr)[1])

R> N <- 1000  # how many operations
R> n <- 999    # matrix dimension
R> cutoff <- 0.9 # what will be set to 0

R> A <- array( runif(n^2), c(n,n))
R> A[A < cutoff] <- 0
R> S <- somecalltomagicfunctiontogetsparsesformat(A)

Compare timing for different operations on A and S.
Timing

R> timing( A + sqrt(A) )
  [1] 0.058
R> timing( S + sqrt(S) )
  [1] 0.061

R> timing( AtA <- t(A) %*% A )
  [1] 0.467
R> timing( StS <- t(S) %*% S )
  [1] 4.222
R> timing( A[1,2] <- .5 )
[1] 0.007
R> timing( S[1,2] <- .5 )
[1] 0.018

R> timing( A[n,n-1] <- .5 )
[1] 0.001
R> timing( S[n,n-1] <- .5 )
[1] 0.012
Timing

R> timing( xA <- solve(AtA, rep(1,n)) )
[1] 1.116
R> timing( xS <- solve(StS, rep(1,n)) )
[1] 1.51

R> timing( RA <- chol(AtA) )
[1] 0.488
R> timing( RS <- chol(StS) )
[1] 1.504

Is it really worthwhile? What is going on?
Timing

Matrix $S$

Matrix $StS$
Timing

With cutoff 0.99:

R> timing(AtA <- t(A) %*% A )
[1] 0.106
R> timing(StS <- t(S) %*% S )
[1] 0.089

R> timing( RA <- chol(AtA) )
[1] 0.494
R> timing( RS <- chol(StS) )
[1] 0.451
Density of the factor is 41% with fill-in ratio 7.2.
Timing

With cutoff 0.999:

R> timing(AtA <- t(A) %*% A)
[1] 0.059
R> timing(StS <- t(S) %*% S)
[1] 0.002

R> timing(RA <- chol(AtA))
[1] 0.466
R> timing(RS <- chol(StS))
[1] 0.007
Density of the factor is .6% with fill-in ratio 2.3.
Implications

With a new storage format new “algorithms” are required . . .

Is it worthwhile???

Yes!

Especially since

spam: R package for sparse matrix algebra.
What is spam?

- an R package for **sparse** matrix algebra
  - publicly available from CRAN, 0.15-3
  - platform independent and documented

- storage economical and fast
  - uses “old Yale sparse format”
  - most routines are in Fortran, adapted for **spam**
  - balance between readability and overhead
  - flags for “expert” users

- versatile, intuitive and simple
  - wrap an **as.spam()** and go
  - **S4** and **S3** syntax

- situated between **SparseM** and **Matrix**
Representation of Sparse Matrices

`spam` defines a S4 class `spam` containing the vectors: “entries”, “colindices”, “rowpointers” and “dimension”.

```r
R> slotNames( "spam")
[1] "entries"    "colindices"  "rowpointers"  "dimension"

R> getSlots( "spam")
entries   colindices  rowpointers  dimension
"numeric"  "integer"  "integer"  "integer"
```
Representation of Sparse Matrices

R> A
[1,]  1.0  0.1  0.0  0.2  0.3
[2,]  0.6  2.0  0.0  0.5  0.0
[3,]  0.0  0.0  3.0  0.0  0.6
[4,]  0.7  0.8  0.0  4.0  0.0
[5,]  0.9  0.0  1.0  0.0  5.0
Class 'spam'
R> slotNames(A)
[1] "entries"  "colindices"  "rowpointers"  "dimension"
R> A@entries
 [1] 1.0 0.1 0.2 0.3 0.6 2.0 0.5 3.0 0.6 0.7 0.8 4.0 0.9 1.0 5.0
R> A@colindices
 [1] 1 2 4 5 1 2 4 3 5 1 2 4 1 3 5
R> A@rowpointers
 [1] 1 5 8 10 13 16
R> A@dimension
 [1] 5 5
Creating Sparse Matrices

Similar coercion techniques as with \texttt{matrix}:

- \texttt{spam(...)}
- \texttt{as.spam(...)}

Special functions:

- \texttt{diag.spam(...)}
- \texttt{nearest.dist(...)}
Methods for spam

- Similar behavior as with matrices
  plot; dim; determinant; %*%; +; ...

- Slightly enhanced behavior
  print; dim<--; chol;

- Specific behavior
  Math; Math2; Summary; ...

- New methods
  display; ordering;
Create Covariance Matrices

Covariance matrix:

nearest.dist and applying a covariance function:

R> C <- nearest.dist(x, diag=TRUE, upper=NULL)
R> C@entries <- Wendland( C@entries, dim=2, k=1)

Precision matrix (GMRF):
— regular grids: nearest.dist with different cutoffs

R> diag.spam(n) +
+ (b1-b2) * nearest.dist(x, delta=1, upper=NULL) +
+ b2 * nearest.dist(x, delta=sqrt(2), upper=NULL)

— irregular grids: using incidence list and spam

R> incidence <- list( i=..., j=..., values)
R> C <- spam( incidence, n, n)
A key feature of spam is to solve efficiently linear systems.

To solve the system $Ax = b$, we

- perform a Cholesky factorisation $A = U^T U$
- solve two triangular systems $U^T z = b$ and $U x = z$

But we need to "ensure" that $U$ is as sparse as possible!

Permute the rows and columns of $A$: $P^T A P = U^T U$. 
Some technical details about a Cholesky decomposition:

<table>
<thead>
<tr>
<th></th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Determine permutation and permute the input matrix $A$ to obtain $P^TAP$</td>
</tr>
<tr>
<td>2</td>
<td>Symbolic factorization: the sparsity structure of $U$ is constructed</td>
</tr>
<tr>
<td>3</td>
<td>Numeric factorization: the elements of $U$ are computed</td>
</tr>
</tbody>
</table>
spam knows Cholesky!

— Several methods to construct permutation matrices $P$
— update to perform only ‘partial’ Cholesky factors
— Flags for avoiding sanity checks
Cholesky

Cholesky factor with MMD

Density: 1.5%, fill-in: 4.7

RCM ordering

Density: 2.7%, fill-in: 8.1
Time and memory usage for 101 Cholesky factorizations (solid) and one factorization and 100 updates (dashed) of a precision matrix from different sizes $L$ of regular $L \times L$ grids with a second order neighbor structure.

(The precision matrix from $L = 200$ has $L^4 = 1.6 \cdot 10^9$ elements). See also demo("article-jss").
Gain of time and memory usage with different options and arguments in the case of a second order neighbor structure of a regular $50 \times 50$ grid and of the US counties. The time and memory usage for the generic call `chol` are 6.2 seconds, 174.5 Mbytes and 15.1 seconds, 416.6 Mbytes, respectively.

<table>
<thead>
<tr>
<th>Options or arguments</th>
<th>Regular grid time</th>
<th>Regular grid memory</th>
<th>US counties time</th>
<th>US counties memory</th>
</tr>
</thead>
<tbody>
<tr>
<td>Using the specific call <code>chol.spam</code></td>
<td>1.001</td>
<td>0.992</td>
<td>0.954</td>
<td>1.004</td>
</tr>
<tr>
<td>Option <code>safemode=c(FALSE,FALSE,FALSE)</code></td>
<td>0.961</td>
<td>1.002</td>
<td>0.988</td>
<td>0.997</td>
</tr>
<tr>
<td>Option <code>chol symmetrycheck=FALSE</code></td>
<td>0.568</td>
<td>0.524</td>
<td>0.646</td>
<td>0.493</td>
</tr>
<tr>
<td>Passing memory=list(<code>nnzR=...</code>, <code>nnzcolindices=...</code>)</td>
<td>0.969</td>
<td>0.979</td>
<td>0.928</td>
<td>0.972</td>
</tr>
<tr>
<td>All of the above</td>
<td>0.561</td>
<td>0.508</td>
<td>0.618</td>
<td>0.490</td>
</tr>
<tr>
<td>All of the above and passing <code>pivot=...</code> to <code>chol.spam</code></td>
<td>0.542</td>
<td>0.528</td>
<td>0.572</td>
<td>0.496</td>
</tr>
<tr>
<td>All of the above and option <code>chol pivot check=FALSE</code></td>
<td>0.510</td>
<td>0.511</td>
<td>0.557</td>
<td>0.489</td>
</tr>
<tr>
<td>Numeric update only using <code>update</code></td>
<td>0.132</td>
<td>0.070</td>
<td>0.170</td>
<td>0.063</td>
</tr>
</tbody>
</table>
BTW:
efficient Cholesky factorization $\iff$ efficient determinant calculation:

$$\det(C) = \det(U^T) \det(U) = \prod_{i=1}^{n} U_{ii}^2$$
Options in spam

For "experts", flags to speed up the code:

R> powerboost() # in spam_0.15-4

R> noquote( format( spam.options()) )

<table>
<thead>
<tr>
<th>eps</th>
<th>drop</th>
<th>printsize</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.220446e-16</td>
<td>FALSE</td>
<td>100</td>
</tr>
<tr>
<td>imagesize</td>
<td>trivalues</td>
<td>cex</td>
</tr>
<tr>
<td>10000</td>
<td>FALSE</td>
<td>1200</td>
</tr>
<tr>
<td>safemode</td>
<td>dopivoting</td>
<td>cholsymmetrycheck</td>
</tr>
<tr>
<td>TRUE, TRUE, TRUE</td>
<td>TRUE</td>
<td>TRUE</td>
</tr>
<tr>
<td>cholpivotcheck</td>
<td>cholupdatesingular</td>
<td>cholincreasefactor</td>
</tr>
<tr>
<td>TRUE</td>
<td>warning</td>
<td>1.25, 1.25</td>
</tr>
<tr>
<td>nearestdistincreasefactor</td>
<td>nearestdistnnz</td>
<td></td>
</tr>
<tr>
<td>1.25</td>
<td>160000, 400</td>
<td></td>
</tr>
</tbody>
</table>
Limits of spam

What can spam not do (yet)?

- LU decompositions
- SVD/eigendecompositions
- Non-double elements
- ...

But, please, comments to rfurrer@mines.edu!
Sparse Matrices in Statistics

Where do large matrices occur?

- Location matrices
- Design matrices
- Covariance matrices
- Precision matrices
Sparse Matrices in Statistics

- Covariance matrices:
  - Compactly supported covariance functions
  - Tapering

- Precision matrices:
  - (Gaussian) Markov random fields
  - (Tapering???)

We have symmetric positive definite (spd) matrices.
Motivation

Precipitation anomaly along 40° lat.
Motivation

Precipitation anomaly along 40° lat.

Tapering

Anomaly

Longitude

Boulder
Lafayette (15km)
Woodrow (156km)
Indianapolis (1632km)
For an isotropic and stationary process with Matérn covariance $C_0(h)$, find a taper $C_\theta(h)$, such that kriging with the product $C_0(h) C_\theta(h)$ is asymptotically optimal.
Objective

For an isotropic and stationary process with Matérn covariance $C_0(h)$, find a taper $C_\theta(h)$, such that kriging with the product $C_0(h)C_\theta(h)$ is asymptotically optimal.

\[
\frac{\text{MSE}(x^*, C_0C_\theta)}{\text{MSE}(x^*, C_0)} \to 1 \\
\frac{\rho(x^*, C_0C_\theta)}{\text{MSE}(x^*, C_0)} \to \gamma
\]

\[
\rho(x^*, C) = C(0) - c^T C^{-1} c^* \]
Misspecified Covariances

In a series of (Annals) papers, Stein gives asymptotic results for misspecified covariances.

Under appropriate conditions, tapering is a form of misspecification.

The taper has to be

- as differentiable at the origin as the original covariance
- more differentiable throughout the domain than at the origin
Tapering

Tapering is an (asymptotically and computationally) efficient technique to create sparse covariance matrices.

Taper range can be justified by computing resources. However, 20–30 locations within the taper range is often sufficient.

“Classical” tapers are:

- spherical: $C_\theta(h) = \left(1 - \frac{|h|}{\theta}\right)^2 + \left(1 + \frac{|h|}{\theta}\right)$

- Wendland-type: $C_\theta(h) = \left(1 - \frac{|h|}{\theta}\right)^{\ell + k} \cdot \text{polynomial in } \frac{|h|}{\theta} \text{ of deg } k$
Positive Definite Matrices

A (large) covariance (often) appears in:

- drawing from a multivariate normal distribution
- calculating/maximizing the (log-)likelihood
- linear/quadratic discrimination analysis
- PCA, EOF, . . .

But all boils down to solving a linear system and possibly calculating the determinant . . .

'Sparse PCA' is sparse in a different sense . . .
Sparse Matrices and fields

- fields is not bound to a specific sparse matrix format
- All heavy lifting is done in mKrig or Krig.engine.fixed
- For a specific sparse format, requires the methods: chol, backsolve, forwardsolve and diag as well as elementary matrix operations need to exist
- If available uses operators to handle diagonal matrices quickly

⇝ The covariance matrix has to stem from particular class.
fields uses spam as default package!
Example mKrig

With appropriate covariance function:

```r
R> x <- USprecip[ precipsubset, 1:2] # locations
R> Y <- USprecip[ precipsubset, 4] # anomaly
R> out <- mKrig(x,Y, m=1, cov.function="wendland.cov",theta=1.5)

R> out.p <- predict.surface( out, nx=220, ny=120)
R> surface(out.p, type='I')
R> US(add=T)
R> points(x,pch='.’)
```
Example Krig

R> out <- Krig( x,Y, m=1, cov.function="wendland.cov", theta=1.5,  
                lambda=0)

R> out.q <- predict.surface( out, nx=220, ny=120)

R> sum( ( out.q$z-out.p$z)^2, na.rm=T)
[1] 1.616783e-20

Krig/predict is slower (here 2.1/3.7 vs 10.4/3.9 seconds).
wendland.cov (based on Wendland) produces a spam matrix. All matrix functions are appropriately overloaded ...

To create sparse covariance matrices based on other covariance functions, use wendland.cov as skeleton.
Tapering

Tapering can be performed with `stationary.taper.cov`. Arguments are (selection):

- **Covariance** = "Exponential"
- **Taper** = "Wendland"
- **Taper.args** = NULL: arguments for the taper function
- **Dist.args** = NULL: arguments passed to `nearest.dist`
- **...** : arguments passed to covariance function

All arguments can also be passed from `mKrig/Krig`
Tapering

Compare the predicted surfaces without and with tapering:

R> out1 <- mKrig( xr,Yr, m=1, theta=1.5) 
R> out1.p <- predict.surface( out1, nx=220, ny=120) 

R> out2 <- mKrig( xr,Yr, m=1, theta=1.5, 
+ cov.function="stationary.taper.cov", 
+ Taper.args = list(k=0, theta=3))) 
R> out2.p <- predict.surface( out2, nx=220, ny=120))

(timing yields 4/22 and 1/9 seconds)
Tapering

Exponential covariance and with tapering
Asides . . .

The following arguments of mKrig/Krig are linked to spam:

- **Dist.args**: arguments passed to nearest.dist
- **chol.args**: arguments passed to chol

Use their help for fine tuning.

**predict.se.Krig, predict.surface.se.Krig** are very inefficient because `nrow(x)` equations need to be solved.
How Big is Big?

Upper limit to create a large matrix is the minimum of:

1. available memory (machine and OS/shell dependent)
   Error: 'cannot allocate vector of size'
2. addressing capacity \((2^{31} - 1)\)
   Error: 'cannot allocate vector of length'

However, R is based on passing by value, calls create local copies (often 3–4 times the space of the object is used).

R> help("Memory-limits")
And Beyond?

Parallelization:
nws, snow, Rmpi, ...

Memory “Outsourcing”:
Matrices are not (entirely) kept in memory:
ff, filehash, biglm, ...

(S+ has the library BufferedMatrix)
