Exploiting sparsity in model matrices

Douglas Bates and Martin Maechler

Department of Statistics
University of Wisconsin – Madison   U.S.A.

Seminar für Statistik
ETH Zurich   Switzerland
(bates|maechler)@R-project.org   (R-Core)

DSC2009, Copenhagen
July 14, 2009

Outline

1. Model matrices with many columns
2. Applications to linear mixed models
3. The penalized least squares problem
4. Using the sparse Cholesky for mixed models
5. Evaluating the likelihood
6. More general model forms
7. Who's the best liked prof at ETH?

Outline

1. Model matrices with many columns

2. Applications to linear mixed models

3. The penalized least squares problem

4. Using the sparse Cholesky for mixed models

5. Evaluating the likelihood

6. More general model forms

7. Who's the best liked prof at ETH?

Outline

1. Model matrices with many columns
2. Applications to linear mixed models
3. The penalized least squares problem
4. Using the sparse Cholesky for mixed models
5. Evaluating the likelihood
6. More general model forms
7. Who's the best liked prof at ETH?

Outline

1. Model matrices with many columns

2. Applications to linear mixed models

3. The penalized least squares problem

4. Using the sparse Cholesky for mixed models

5. Evaluating the likelihood

6. More general model forms

7. Who's the best liked prof at ETH?

Outline

1. Model matrices with many columns
2. Applications to linear mixed models
3. The penalized least squares problem
4. Using the sparse Cholesky for mixed models
5. Evaluating the likelihood
6. More general model forms
7. Who's the best liked prof at ETH?

Outline

1. Model matrices with many columns

2. Applications to linear mixed models

3. The penalized least squares problem

4. Using the sparse Cholesky for mixed models

5. Evaluating the likelihood

6. More general model forms

7. Who's the best liked prof at ETH?

Outline

1. Model matrices with many columns

2. Applications to linear mixed models

3. The penalized least squares problem

4. Using the sparse Cholesky for mixed models

5. Evaluating the likelihood

6. More general model forms

7. Who's the best liked prof at ETH?

Outline

1. Model matrices with many columns

2. Applications to linear mixed models

3. The penalized least squares problem

4. Using the sparse Cholesky for mixed models

5. Evaluating the likelihood

6. More general model forms

7. Who's the best liked prof at ETH?

In case it all starts to blur

- Model matrices with many columns typically have some degree of sparsity.
- Regularization is important in combination with many columns. Updating a factorization is much more efficient when optimizing a regularization parameter.
- The complexity of the factorization depends on the order of the columns. Need to consider carefully the parameterization (contrasts) and the order of terms.
- Mixed-effects models are regularization problems that benefit greatly from sparse matrix methods. They are implicit in `lme4`.
- A new function `sparse.model.matrix()` is available in the `Matrix` package for non-implicit sparse model matrix construction.
- These slides are available at `http://Matrix.R-forge.R-project.org/slides/`.

Model matrices and sparsity

- In statistical models the effects of the covariates on the response are often expressed, directly or indirectly, through *model matrices*. A common idiom in a model fitting function using a `formula` argument is a call to `model.frame()` followed by a call to `model.matrix()`.
- Many users feel frustrated that R does not transparently handle very large model matrices, failing to realize that a naive decomposition of an $n \times p$ dense model matrix requires $O(np^2)$ flops. Large values of p are thus particularly problematic.
- Frequently, large values of p are a consequence of incorporating factors with a large number of levels in the model. A factor with k levels generates at least $k - 1$ columns as do any interactions with such a factor.
- The model matrix columns are generated from the indicator columns for the factor, which are very sparse. The greater the number of levels, the more sparse the indicators become.

Sparse model matrices and regularization

- As stated at useR!2009, large, sparse model matrices usually require some amount of regularization for computationally feasible evaluation of coefficients and fitted values.
- Frequently the regularization parameter(s) are chosen to optimize a criterion, requiring evaluation of the criterion for many different trial values of the regularization parameter(s).
- Usually the repeated evaluations of the criterion require decomposition of a matrix with a constant structure (including the positions of the non-zeros) and varying numeric values.
- The sparse Cholesky factorization is ideally suited to problems requiring many evaluations of a decomposition of a matrix with constant structure and varying numeric values.

The sparse Cholesky factorization

- The `Matrix` package for R provides sparse matrix methods, including the sparse Cholesky, by interfacing to Tim Davis' CHOLMOD library of C functions.
- This C library provides separate functions for the symbolic factorization, including determining a *fill-reducing permutation*, and the numeric factorization.
- The symbolic factorization determines the positions of the non-zeros in the result. The numeric factorization simply evaluates the numeric values. Generally it is much faster than the symbolic factorization.
- There are many beautiful mathematical results associated with sparse matrix operations. See Tim Davis' 2007 SIAM book for some of these results.

Variations of the sparse Cholesky

- In the `Matrix` package we use the formulation from the CHOLMOD C library. Sparse matrices may be entered in the triplet formulation but operations are usually performed on the *compressed-column representation* (the `CsparseMatrix` class).
- If $\boldsymbol{A}$ is a positive-definite symmetric sparse matrix, the sparse Cholesky factorization consists of a permutation matrix $\boldsymbol{P}$ and a lower triangular matrix $\boldsymbol{L}$ such that

  $$\boldsymbol{L}\boldsymbol{L}^{\mathsf{T}} = \boldsymbol{P}\boldsymbol{A}\boldsymbol{P}^{\mathsf{T}}.$$

  Note that $\boldsymbol{L}$ is the left factor (statisticians often consider the right factor, $\boldsymbol{R} = \boldsymbol{L}^{\mathsf{T}}$). The permutation $\boldsymbol{P}$ is stored (as a vector) within the factorization.
- There are two variations: the *LDL* factorization, where the lhs is $\boldsymbol{L}\boldsymbol{D}\boldsymbol{L}^{\mathsf{T}}$ ($\boldsymbol{L}$ unit lower triangular; $\boldsymbol{D}$ diagonal), and a *supernodal* $\boldsymbol{L}\boldsymbol{L}^{\mathsf{T}}$ decomposition, which is a sparse/dense hybrid that collapses columns with similar structure to a “supernode” of the graph representation.

Outline

1. Model matrices with many columns
2. Applications to linear mixed models
3. The penalized least squares problem
4. Using the sparse Cholesky for mixed models
5. Evaluating the likelihood
6. More general model forms
7. Who's the best liked prof at ETH?

Definition of linear mixed models

- A linear mixed model consists of two random variables: the n-dimensional response, $\boldsymbol{\mathcal{Y}}$, and the q-dimensional random effects, $\boldsymbol{\mathcal{B}}$. We observe the value, $\boldsymbol{y}$, of $\boldsymbol{\mathcal{Y}}$; we do not observe the value of $\boldsymbol{B}$.
- The probability model defines one conditional and one unconditional distribution

$$(\boldsymbol{\mathcal{Y}}|\boldsymbol{\mathcal{B}} = \boldsymbol{b}) \sim \mathcal{N}\left(\boldsymbol{Z}\boldsymbol{b} + \boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_n\right), \quad \boldsymbol{\mathcal{B}} \sim \mathcal{N}\left(\boldsymbol{0}, \boldsymbol{\Sigma}_\theta\right),$$

  which depend on parameters $\boldsymbol{\beta}$, $\boldsymbol{\theta}$ and σ.
- Although the dimension of $\boldsymbol{\Sigma}_\theta$ can be huge, the dimension of the *variance-component parameter vector*, $\boldsymbol{\theta}$, is usually very small.
- The model specification determines the $n \times q$ model matrix $\boldsymbol{Z}$ (generated from indicator columns and typically very sparse), the $n \times p$ model matrix $\boldsymbol{X}$, and the way in which $\boldsymbol{\theta}$ generates $\boldsymbol{\Sigma}_\theta$.

Properties of Σ_θ; generating it

- Because it is a variance-covariance matrix, the $q \times q$ $\boldsymbol{\Sigma}_\theta$ must be symmetric and *positive semi-definite*, which means, in effect, that it has a "square root" — there must be another matrix that, when multiplied by its transpose, gives $\boldsymbol{\Sigma}_\theta$.
- We never really form $\boldsymbol{\Sigma}_\theta$; we always work with the *relative covariance factor*, $\boldsymbol{\Lambda}_\theta$, defined so that

$$\boldsymbol{\Sigma}_\theta = \sigma^2 \boldsymbol{\Lambda}_\theta \boldsymbol{\Lambda}_\theta^\mathsf{T}$$

  where σ^2 is the same variance parameter as in $(\boldsymbol{\mathcal{Y}}|\boldsymbol{\mathcal{B}} = \boldsymbol{b})$.
- We also work with a q-dimensional "spherical" or "unit" random-effects vector, $\boldsymbol{\mathcal{U}}$, such that

$$\boldsymbol{\mathcal{U}} \sim \mathcal{N}\left(\mathbf{0}, \sigma^2 \boldsymbol{I}_q\right), \; \boldsymbol{\mathcal{B}} = \boldsymbol{\Lambda}_\theta \boldsymbol{\mathcal{U}} \Rightarrow \mathsf{Var}(\boldsymbol{\mathcal{B}}) = \sigma^2 \boldsymbol{\Lambda}_\theta \boldsymbol{\Lambda}_\theta^\mathsf{T} = \boldsymbol{\Sigma}_\theta.$$

- The linear predictor expression becomes

$$\boldsymbol{Z b} + \boldsymbol{X \beta} = \boldsymbol{Z \Lambda}_\theta \, \boldsymbol{u} + \boldsymbol{X \beta} = \boldsymbol{U}_\theta \, \boldsymbol{u} + \boldsymbol{X \beta}$$

  where $\boldsymbol{U}_\theta = \boldsymbol{Z \Lambda}_\theta$.

The conditional mode, $\tilde{\boldsymbol{u}}_{\theta,\beta}$

- Although the probability model is defined from $(\boldsymbol{\mathcal{Y}}|\boldsymbol{\mathcal{U}}=\boldsymbol{u})$, we observe $\boldsymbol{y}$, not $\boldsymbol{u}$ (or $\boldsymbol{b}$) so we want to work with the other conditional distribution, $(\boldsymbol{\mathcal{U}}|\boldsymbol{\mathcal{Y}}=\boldsymbol{y})$.
- The joint distribution of $\boldsymbol{\mathcal{Y}}$ and $\boldsymbol{\mathcal{U}}$ is Gaussian with density

$$\begin{aligned} f_{\boldsymbol{\mathcal{Y}},\boldsymbol{\mathcal{U}}}(\boldsymbol{y},\boldsymbol{u}) &= f_{\boldsymbol{\mathcal{Y}}|\boldsymbol{\mathcal{U}}}(\boldsymbol{y}|\boldsymbol{u})\, f_{\boldsymbol{\mathcal{U}}}(\boldsymbol{u}) \\ &= \frac{\exp(-\frac{1}{2\sigma^2}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{U}_\theta\,\boldsymbol{u}\|^2)}{(2\pi\sigma^2)^{n/2}}\;\frac{\exp(-\frac{1}{2\sigma^2}\|\boldsymbol{u}\|^2)}{(2\pi\sigma^2)^{q/2}} \\ &= \frac{\exp(-\left[\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{U}_\theta\,\boldsymbol{u}\|^2+\|\boldsymbol{u}\|^2\right]/(2\sigma^2))}{(2\pi\sigma^2)^{(n+q)/2}} \end{aligned}$$

- The mode, $\tilde{\boldsymbol{u}}_{\theta,\beta}$, of the conditional distribution $(\boldsymbol{\mathcal{U}}|\boldsymbol{\mathcal{Y}}=\boldsymbol{y})$ (also the mean in this case) is

$$\tilde{\boldsymbol{u}}_{\theta,\beta} = \arg\min_{\boldsymbol{u}}\left[\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{U}_\theta\,\boldsymbol{u}\|^2+\|\boldsymbol{u}\|^2\right]$$

Outline

1 Model matrices with many columns

2 Applications to linear mixed models

3 The penalized least squares problem

4 Using the sparse Cholesky for mixed models

5 Evaluating the likelihood

6 More general model forms

7 Who's the best liked prof at ETH?

Minimizing a penalized sum of squared residuals

- An expression like $\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{U}_\theta\,\boldsymbol{u}\|^2 + \|\boldsymbol{u}\|^2$ is called a *penalized sum of squared residuals* because $\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{U}_\theta\,\boldsymbol{u}\|^2$ is a sum of squared residuals and $\|\boldsymbol{u}\|^2$ is a penalty on the size of the vector $\boldsymbol{u}$.
- Determining $\tilde{\boldsymbol{u}}_{\theta,\beta}$ as the minimizer of this expression is a *penalized least squares* (PLS) problem. In this case it is a *penalized linear least squares problem* that we can solve directly (i.e. without iterating).
- One way to determine the solution is to rephrase it as a linear least squares problem for an extended residual vector

$$\tilde{\boldsymbol{u}}_{\theta,\beta} = \arg\min_{\boldsymbol{u}} \left\| \begin{bmatrix} \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} \\ \boldsymbol{0} \end{bmatrix} - \begin{bmatrix} \boldsymbol{U}_\theta \\ \boldsymbol{I}_q \end{bmatrix} \boldsymbol{u} \right\|^2$$

  This is sometimes called a *pseudo-data* approach because we create the effect of the penalty term, $\|\boldsymbol{u}\|^2$, by adding “pseudo-observations” to $\boldsymbol{y}$ and to the predictor.

Solving the linear PLS problem

- The conditional mean satisfies the equations

$$(\boldsymbol{U}_\theta^\mathsf{T}\boldsymbol{U}_\theta + \boldsymbol{I}_q)\tilde{\boldsymbol{u}}_{\theta,\beta} = \boldsymbol{U}_\theta^\mathsf{T}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$$

- This would be interesting but not very important were it not for the fact that we actually can solve that system for $\tilde{\boldsymbol{u}}_{\theta,\beta}$ even when its dimension, q, is very, very large.
- Recall that $\boldsymbol{U}_\theta = \boldsymbol{Z}\boldsymbol{\Lambda}_\theta$. Because $\boldsymbol{Z}$ is generated from indicator columns for the grouping factors, it is sparse. $\boldsymbol{U}_\theta$ is also very sparse.
- The fill-reducing permutation, $\boldsymbol{P}$, and the structure of the Cholesky factor, $\boldsymbol{L}$, are determined from $\boldsymbol{U}_{\theta^{(0)}}$ where $\boldsymbol{\theta}^{(0)}$ is the starting value. For subsequent values of $\boldsymbol{\theta}$ the update of the factor $\boldsymbol{L}_\theta$ satisfying

$$\boldsymbol{L}_\theta\boldsymbol{L}_\theta^\mathsf{T} = \boldsymbol{P}\left(\boldsymbol{U}_\theta^\mathsf{T}\boldsymbol{U}_\theta + \boldsymbol{I}_q\right)\boldsymbol{P}^\mathsf{T}$$

  is direct from $\boldsymbol{U}_\theta$. (One of the CHOLMOD functions does the update, including virtually adding a multiple of the identity, from the sparse, rectangular $\boldsymbol{U}_\theta$.) From $\boldsymbol{L}_\theta$ we solve for $\tilde{\boldsymbol{u}}_{\theta,\beta}$.

What do we mean by “large” nowadays?

- Harold Doran recently fit a linear mixed model to the annual achievement test results for the last 4 years in one of the United States. There were $n = 5212017$ observations on a total of $n_1 = 1876788$ students and $n_2 = 47480$ teachers.
- The models had simple, scalar random effects for student and for teacher resulting in $q = 1924268$ (i.e. nearly 2 million!)
- There were a total of $p = 29$ fixed-effects parameters.
- At present Harold needed to fit the model to a subset and only evaluate the conditional means for all the students and teachers but we should be able to get around that limitation and actually fit the model to all these responses and random effects.
- I don’t know of other software that can be used to fit a model this large.

Size of the decomposition for this large model

- Memory usage in such a model is dominated by Cholesky factor, $\boldsymbol{L}(\boldsymbol{\theta})$.
- In this case the `x` slot is itself over 1GB in size (`i` slot $>$ 0.5 GB).
- These are close to an inherent limit on atomic R objects (the range of an index into an atomic object cannot exceed $2^{31}{}_{(=2147'483'648)}$.

```
> str(L)

Formal class 'dCHMsimpl' [package "Matrix"] with 10 slots
  ..@ x       : num [1:174396181] 1.71 2.16 1.4 1.32 2.29 ...
  ..@ p       : int [1:1924269] 0 2 4 5 7 9 10 12 14 15 ...
  ..@ i       : int [1:174396181] 0 2 1 2 2 3 5 4 5 5 ...
  ..@ nz      : int [1:1924268] 2 2 1 2 2 1 2 2 1 2 ...
  ..@ nxt     : int [1:1924270] 1 2 3 4 5 6 7 8 9 10 ...
  ..@ prv     : int [1:1924270] 1924269 0 1 2 3 4 5 6 7 8 ...
  ..@ colcount: int [1:1924268] 2 2 1 2 2 1 2 2 1 2 ...
  ..@ perm    : int [1:1924268] 1922843 1886519 134451 1921046 1893309 1834
  ..@ type    : int [1:4] 2 1 0 1
  ..@ Dim     : int [1:2] 1924268 1924268
```

Outline

1. Model matrices with many columns
2. Applications to linear mixed models
3. The penalized least squares problem
4. Using the sparse Cholesky for mixed models
5. Evaluating the likelihood
6. More general model forms
7. Who's the best liked prof at ETH?

Applications to models with simple, scalar random effects

- For a model with simple, scalar random-effects terms only, the matrix $\boldsymbol{\Sigma}_\theta$ is block-diagonal in k blocks and the ith block is $\sigma_i^2 \boldsymbol{I}_{n_i}$ where n_i is the number of levels in the ith grouping factor.
- The matrix $\boldsymbol{\Lambda}_\theta$ is also block-diagonal with the ith block being $\theta_i \boldsymbol{I}_{n_i}$, where $\theta_i = \sigma_i / \sigma$.
- Given the grouping factors for the model and a value of $\boldsymbol{\theta}$ we produce $\boldsymbol{U}_\theta$ then $\boldsymbol{L}_\theta$, using `Cholesky` the first time then `update`.
- To avoid recalculating we assign

`flist`	a list of the grouping factors
`nlev`	number of levels in each factor
`Zt`	the transpose of the model matrix, $\boldsymbol{Z}$
`theta`	current value of $\boldsymbol{\theta}$
`Lambda`	current $\boldsymbol{\Lambda}_\theta$
`Ut`	transpose of $\boldsymbol{U}_\theta = \boldsymbol{Z}\boldsymbol{\Lambda}_\theta$

Cholesky factor for the Penicillin model

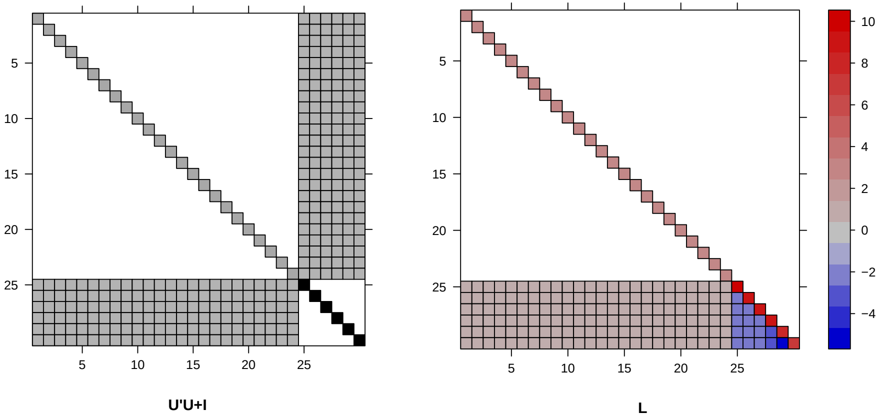
```
> flist <- subset(Penicillin, select = c(plate, sample))
> Zt <- do.call(rBind, lapply(flist, as, "sparseMatrix"))
> (nlev <- sapply(flist, function(f) length(levels(factor(f)))))
```

```
 plate sample
    24      6
```

```
> theta <- c(1.2, 2.1)
> Lambda <- Diagonal(x = rep.int(theta, nlev))
> Ut <- crossprod(Lambda, Zt)
> str(L <- Cholesky(tcrossprod(Ut), LDL = FALSE, Imult = 1))
```

```
Formal class ’dCHMsimpl’ [package "Matrix"] with 10 slots
  ..@ x        : num [1:189] 3.105 0.812 0.812 0.812 0.812 ...
  ..@ p        : int [1:31] 0 7 14 21 28 35 42 49 56 63 ...
  ..@ i        : int [1:189] 0 24 25 26 27 28 29 1 24 25 ...
  ..@ nz       : int [1:30] 7 7 7 7 7 7 7 7 7 7 ...
  ..@ nxt      : int [1:32] 1 2 3 4 5 6 7 8 9 10 ...
  ..@ prv      : int [1:32] 31 0 1 2 3 4 5 6 7 8 ...
  ..@ colcount: int [1:30] 7 7 7 7 7 7 7 7 7 7 ...
  ..@ perm     : int [1:30] 23 22 21 20 19 18 17 16 15 14 ...
  ..@ type     : int [1:4] 2 1 0 1
  ..@ Dim      : int [1:2] 30 30
```

Images of $U^\intercal U + I$ and L



- Note that there are nonzeros in the lower right of L in positions that are zero in the lower triangle of $U^\intercal U + I$. This is described as “fill-in”.

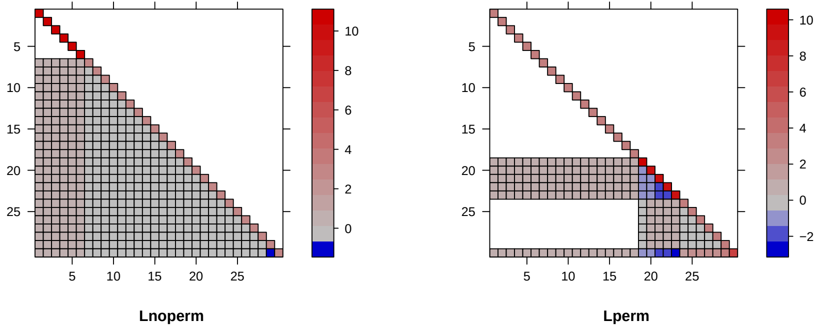
Reversing the order of the grouping factors

- To show the effect of a fill-reducing permutation, we reverse the order of the factors and calculate the Cholesky factor with and without a fill-reducing permutation.
- We evaluate `nnzero` (number of nonzeros) for `L`, from the original factor order, and for `Lnoperm` and `Lperm`, the reversed factor order without and with permutation

```
> Zt <- do.call(rBind, lapply(flist[2:1], as, "sparseMatrix"))
> Lambda <- Diagonal(x = rep.int(theta[2:1], nlev[2:1]))
> Ut <- crossprod(Lambda, Zt)
> Lnoperm <- Cholesky(tcrossprod(Ut), perm = FALSE, LDL = FALSE,
+     Imult = 1)
> Lperm <- Cholesky(tcrossprod(Ut), LDL = FALSE, Imult = 1)
> sapply(lapply(list(L, Lnoperm, Lperm), as, "sparseMatrix"),
+     nnzero)

[1] 189 450 204
```

Images of the reversed factor decompositions



- Without permutation, we get the worst possible fill-in. With a fill-reducing permutation we get much less fill-in but still not as good as the original factor order.
- This is why the permutation is called “fill-reducing”, not “fill-minimizing”. Getting the fill-minimizing permutation in the general case is a very hard problem.

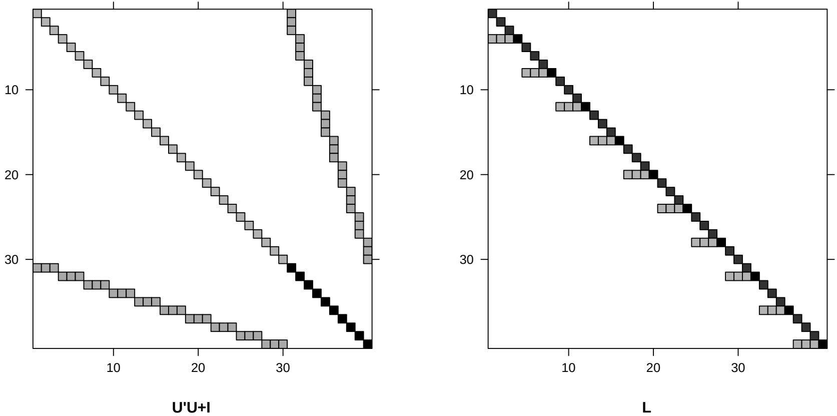
Cholesky factor for the Pastes data

- For the special case of nested grouping factors, such as in the Pastes data, there is no fill-in, regardless of the permutation.
- A permutation is nevertheless evaluated but it is a “post-ordering” that puts the nonzeros near the diagonal.

```
> Zt <- do.call(rBind, lapply(flist <- subset(Pastes,
+     , c(sample, batch)), as, "sparseMatrix"))
> nlev <- sapply(flist, function(f) length(levels(factor(f))))
> theta <- c(0.4, 0.5)
> Lambda <- Diagonal(x = rep.int(theta, nlev))
> Ut <- crossprod(Lambda, Zt)
> L <- Cholesky(tcrossprod(Ut), LDL = FALSE, Imult = 1)
> str(L@perm)
```

```
 int [1:40] 2 1 0 30 5 4 3 31 8 7 ...
```

Image of the factor for the Pastes data



Outline

1. Model matrices with many columns
2. Applications to linear mixed models
3. The penalized least squares problem
4. Using the sparse Cholesky for mixed models
5. **Evaluating the likelihood**
6. More general model forms
7. Who's the best liked prof at ETH?

Evaluating the likelihood for mixed models

- From the joint density, $f_{\mathcal{Y},\mathcal{U}}(\boldsymbol{y},\boldsymbol{u})$, we obtain the likelihood

$$L(\boldsymbol{\theta},\boldsymbol{\beta},\sigma^2|\boldsymbol{y}) = \int f_{\mathcal{Y},\mathcal{U}}(\boldsymbol{y},\boldsymbol{u})\,d\boldsymbol{u}.$$

  The function being integrated is an unnormalized Gaussian density. Its integral can be determined from the value at the mode and the variance-covariance matrix.

- The Cholesky factor, $\boldsymbol{L}_\theta$, allows evaluation of $\tilde{\boldsymbol{u}}_{\theta,\beta}$ from

$$\boldsymbol{P}^{\mathsf{T}}\boldsymbol{L}_\theta\boldsymbol{L}_\theta^{\mathsf{T}}\boldsymbol{P}\tilde{\boldsymbol{u}}_{\theta,\beta} = \boldsymbol{U}_\theta^{\mathsf{T}}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})$$

- The exponent of $f_{\mathcal{Y},\mathcal{U}}(\boldsymbol{y},\boldsymbol{u})$ can now be written

$$\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{U}_\theta\boldsymbol{u}\|^2+\|\boldsymbol{u}\|^2 = r^2(\boldsymbol{\theta},\boldsymbol{\beta})+\|\boldsymbol{L}_\theta^{\mathsf{T}}\boldsymbol{P}(\boldsymbol{u}-\tilde{\boldsymbol{u}}_{\theta,\beta})\|^2.$$

  where $r^2(\boldsymbol{\theta},\boldsymbol{\beta}) = \|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{U}\tilde{\boldsymbol{u}}_{\theta,\beta}\|^2+\|\tilde{\boldsymbol{u}}_{\theta,\beta}\|^2$.

- The first term doesn't depend on $\boldsymbol{u}$ and

$$\int \frac{\exp\left(\frac{-\|\boldsymbol{L}^{\mathsf{T}}\boldsymbol{P}(\boldsymbol{u}-\tilde{\boldsymbol{u}}_{\theta,\beta})\|^2}{2\sigma^2}\right)}{(2\pi\sigma^2)^{q/2}}\,d\boldsymbol{u} = \frac{1}{|\boldsymbol{L}|}$$

Evaluating the likelihood (cont'd)

- As is often the case, it is easiest to write the log-likelihood. On the deviance scale (negative twice the log-likelihood) $\ell(\boldsymbol{\theta}, \boldsymbol{\beta}, \sigma | \boldsymbol{y}) = \log L(\boldsymbol{\theta}, \boldsymbol{\beta}, \sigma | \boldsymbol{y})$ becomes

$$-2\ell(\boldsymbol{\theta}, \boldsymbol{\beta}, \sigma | \boldsymbol{y}) = n \log(2\pi\sigma^2) + \frac{r^2(\boldsymbol{\theta}, \boldsymbol{\beta})}{\sigma^2} + \log(|\boldsymbol{L}_\theta|^2)$$

- We wish to minimize the deviance. Its dependence on σ is straightforward. Given values of the other parameters, we can evaluate the conditional estimate

$$\widehat{\sigma^2}(\boldsymbol{\theta}, \boldsymbol{\beta}) = \frac{r^2(\boldsymbol{\theta}, \boldsymbol{\beta})}{n}$$

  producing the *profiled deviance*

$$-2\tilde{\ell}(\boldsymbol{\theta}, \boldsymbol{\beta} | \boldsymbol{y}) = \log(|\boldsymbol{L}_\theta|^2) + n \left[1 + \log\left(\frac{2\pi r^2(\boldsymbol{\theta}, \boldsymbol{\beta})}{n} \right) \right]$$

- However, an even greater simplification is possible because the deviance depends on $\boldsymbol{\beta}$ only through $r^2(\boldsymbol{\theta}, \boldsymbol{\beta})$.

Profiling the deviance with respect to $\boldsymbol{\beta}$

- Because the deviance depends on $\boldsymbol{\beta}$ only through $r^2(\boldsymbol{\theta}, \boldsymbol{\beta})$ we can obtain the conditional estimate, $\widehat{\boldsymbol{\beta}}_\theta$, by extending the PLS problem to

$$r^2(\boldsymbol{\theta}) = \min_{\boldsymbol{u},\boldsymbol{\beta}} \left[\left\| \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{U}_\theta\, \boldsymbol{u} \right\|^2 + \left\| \boldsymbol{u} \right\|^2 \right]$$

with the solution satisfying the equations

$$\begin{bmatrix} \boldsymbol{U}_\theta^\intercal \boldsymbol{U}_\theta + \boldsymbol{I}_q & \boldsymbol{U}_\theta^\intercal \boldsymbol{X} \\ \boldsymbol{X}^\intercal \boldsymbol{U}_\theta & \boldsymbol{X}^\intercal \boldsymbol{X} \end{bmatrix} \begin{bmatrix} \tilde{\boldsymbol{u}}_\theta \\ \widehat{\boldsymbol{\beta}}_\theta \end{bmatrix} = \begin{bmatrix} \boldsymbol{U}_\theta^\intercal \boldsymbol{y} \\ \boldsymbol{X}^\intercal \boldsymbol{y}. \end{bmatrix}$$

- The profiled deviance, which is a function of $\boldsymbol{\theta}$ only, is

$$-2\tilde{\ell}(\boldsymbol{\theta}) = \log(|\boldsymbol{L}_\theta|^2) + n \left[1 + \log \left(\frac{2\pi r^2(\boldsymbol{\theta})}{n} \right) \right]$$

Solving the extended PLS problem

- For brevity we will no longer show the dependence of matrices and vectors on the parameter $\boldsymbol{\theta}$.
- As before we use the sparse Cholesky decomposition, with $\boldsymbol{L}$ and $\boldsymbol{P}$ satisfying $\boldsymbol{L}\boldsymbol{L}^\intercal = \boldsymbol{P}(\boldsymbol{U}^\intercal\boldsymbol{U} + \boldsymbol{I})\boldsymbol{P}^\intercal$ and $\boldsymbol{c}_{\boldsymbol{u}}$, the solution to $\boldsymbol{L}\boldsymbol{c}_{\boldsymbol{u}} = \boldsymbol{P}\boldsymbol{U}^\intercal\boldsymbol{y}$.
- We extend the decomposition with the $q \times p$ matrix $\boldsymbol{R}_{ZX}$, the upper triangular $p \times p$ matrix $\boldsymbol{R}_X$, and the p-vector $\boldsymbol{c}_{\boldsymbol{\beta}}$ satisfying

$$\begin{aligned}
\boldsymbol{L}\boldsymbol{R}_{ZX} &= \boldsymbol{P}\boldsymbol{U}^\intercal\boldsymbol{X}\\
\boldsymbol{R}_X^\intercal\boldsymbol{R}_X &= \boldsymbol{X}^\intercal\boldsymbol{X} - \boldsymbol{R}_{ZX}^\intercal\boldsymbol{R}_{ZX}\\
\boldsymbol{R}_X^\intercal\boldsymbol{c}_{\boldsymbol{\beta}} &= \boldsymbol{X}^\intercal\boldsymbol{y} - \boldsymbol{R}_{ZX}^\intercal\boldsymbol{c}_{\boldsymbol{u}}
\end{aligned}$$

so that

$$\begin{bmatrix}\boldsymbol{P}^\intercal\boldsymbol{L} & \boldsymbol{0}\\ \boldsymbol{R}_{ZX}^\intercal & \boldsymbol{R}_X^\intercal\end{bmatrix}\begin{bmatrix}\boldsymbol{L}^\intercal\boldsymbol{P} & \boldsymbol{R}_{ZX}\\ \boldsymbol{0} & \boldsymbol{R}_X\end{bmatrix} = \begin{bmatrix}\boldsymbol{U}^\intercal\boldsymbol{U}+\boldsymbol{I} & \boldsymbol{U}^\intercal\boldsymbol{X}\\ \boldsymbol{X}^\intercal\boldsymbol{U} & \boldsymbol{X}^\intercal\boldsymbol{X}\end{bmatrix}.$$

Solving the extended PLS problem (cont'd)

- Finally we solve

$$\boldsymbol{R}_X\widehat{\boldsymbol{\beta}}_\theta = \boldsymbol{c}_{\boldsymbol{\beta}}$$

$$\boldsymbol{L}^\intercal\boldsymbol{P}\tilde{\boldsymbol{u}}_{\theta,\beta} = \boldsymbol{c}_{\boldsymbol{u}} - \boldsymbol{R}_{ZX}\widehat{\boldsymbol{\beta}}_\theta$$

- The profiled REML criterion also can be expressed simply. The criterion is

$$L_R(\boldsymbol{\theta},\sigma^2|\boldsymbol{y}) = \int L(\boldsymbol{\theta},\boldsymbol{\beta},\sigma^2|\boldsymbol{y})\,d\boldsymbol{\beta}$$

  The same change-of-variable technique for evaluating the integral w.r.t. $\boldsymbol{u}$ as $1/\operatorname{abs}(|\boldsymbol{L}|)$ produces $1/\operatorname{abs}(|\boldsymbol{R}_X|)$ here and removes $(2\pi\sigma^2)^{p/2}$ from the denominator. On the deviance scale, the profiled REML criterion is

$$-2\tilde{\ell}_R(\boldsymbol{\theta}) = \log(|\boldsymbol{L}|^2) + \log(|\boldsymbol{R}_x|^2) + (n-p)\left[1 + \log\left(\frac{2\pi r^2(\boldsymbol{\theta})}{n-p}\right)\right]$$

- These calculations can be expressed in a few lines of R code. Assume the environment of `setPars()` contains `y`, `X`, `Zt`, `REML`, `L`, `nlev` and `XtX` ($\boldsymbol{X}^\intercal\boldsymbol{X}$).

Code for evaluating the profiled deviance

```
setPars <- function(theta) {
  stopifnot(is.numeric(theta), length(theta)==length(nlev))
  Ut <- crossprod(Diagonal(x=rep.int(theta,nlev)),Zt)
  L <- update(L, Ut, mult = 1)
  cu <- solve(L, solve(L, Ut %*% y, sys = "P"), sys = "L")
  RZX <- solve(L, solve(L, Ut %*% X, sys = "P"), sys = "L")
  RX <- chol(XtX - crossprod(RZX))
  cb <- solve(t(RX),crossprod(X,y)- crossprod(RZX, cu))
  beta <- solve(RX, cb)
  u <- solve(L,solve(L,cu - RZX %*% beta, sys="Lt"), sys="Pt")
  fitted <- as.vector(crossprod(Ut, u) + X %*% beta)
  prss <- sum(c(y - fitted, as.vector(u))^2)
  n <- length(fitted);  p <- ncol(RX)
  if (REML) return(determinant(L)$mod +
                   2 * determinant(RX)$mod +
                   (n-p) * (1+log(2*pi*prss/(n-p))))
  determinant(L)$mod + n * (1 + log(2*pi*prss/n))
}
```

How lmer works

- Essentially `lmer` takes its arguments and creates an environment containing the model matrices, the response and the Cholesky factor. The optimization of the profiled deviance or the profiled REML criterion happens within this environment.
- The creation of $\mathbf{\Lambda}_\theta$ is somewhat more complex for models with vector-valued random effects but not excessively so.
- Some care is taken to avoid allocating storage for large objects during each function evaluation. Many of the objects created in `profDev` are updated in place within `lmer`.
- Once the optimizer, `nlminb`, has converged some additional information for the summary is calculated.

Outline

1 Model matrices with many columns

2 Applications to linear mixed models

3 The penalized least squares problem

4 Using the sparse Cholesky for mixed models

5 Evaluating the likelihood

6 More general model forms

7 Who's the best liked prof at ETH?

Nonlinear and generalized linear mixed models

- In a nonlinear mixed model (NMM) the conditional distribution, $(\boldsymbol{\mathcal{Y}}|\boldsymbol{\mathcal{U}} = \boldsymbol{u})$, is Gaussian but its mean depends on the linear predictor, $\boldsymbol{U}_\theta\, \boldsymbol{u} + \boldsymbol{X}\boldsymbol{\beta}$, through a nonlinear model function.
- The conditional mode, $\tilde{\boldsymbol{u}}_{\theta,\beta}$, is the solution to a *penalized nonlinear least squares* problem. The Laplace approximation to the profiled deviance is

  $$-2\tilde{\ell}(\boldsymbol{\theta}, \boldsymbol{\beta}|\boldsymbol{y}) = \log(|\boldsymbol{L}_{\theta,\beta}|^2) + n\left[1 + \log\left(\frac{2\pi r^2(\boldsymbol{\theta}, \boldsymbol{\beta})}{n}\right)\right]$$

  where, as before, $r^2(\boldsymbol{\theta}, \boldsymbol{\beta})$ is the minimum penalized residual sum of squares. The matrix $\boldsymbol{U}_{\theta,\beta}$ determining $\boldsymbol{L}_{\theta,\beta}$ is $\frac{d\boldsymbol{\mu}}{d\boldsymbol{u}^\intercal}$
- For generalized linear mixed models (GLMMs) the penalized least squares problem to determine $\tilde{\boldsymbol{u}}_{\theta,\beta}$ is replaced by a penalized iteratively reweighted least squares problem.
- All of the PLS problems require updating the decomposition for revised numeric values.

Outline

1. Model matrices with many columns
2. Applications to linear mixed models
3. The penalized least squares problem
4. Using the sparse Cholesky for mixed models
5. Evaluating the likelihood
6. More general model forms
7. Who’s the best liked prof at ETH?

Who's the best liked prof at ETH?

- Private donation for encouraging excellent teaching at ETH
- Student union of ETH Zurich organizes survey to award prizes: Best lecturer — of ETH, and of each of the 15 departments.
- Smart Web-interface for survey: Each student sees the names of his/her professors from the last 4 semesters and all the lectures that applied.
- ratings in $\{1, 2, 3, 4, 5\}$.
- high response rate

Who's the best prof — data

```
> str(d.eth)
```

```
'data.frame': 73421 obs. of  7 variables:
 $ s      : Factor w/ 2972 levels "1","2","3","4",..: 1 1 1 1 2 2 3 3 3 3 .
 $ d      : Factor w/ 1128 levels "1","6","7","8",..: 525 560 832 1068 62 4
 $ studage: Ord.factor w/ 4 levels "2"<"4"<"6"<"8": 1 1 1 1 1 1 1 1 1 1 ...
 $ lectage: Ord.factor w/ 6 levels "1"<"2"<"3"<"4"<..: 2 1 2 2 1 1 1 1 1 1
 $ service: Factor w/ 2 levels "0","1": 1 2 1 2 1 1 2 1 1 1 ...
 $ dept   : Factor w/ 15 levels "5","12","6","11",..: 15 5 15 12 2 2 14 3 3
 $ y      : int  5 2 5 3 2 4 4 5 5 4 ...
```

Modelling the ETH teacher ratings

Model: The rating depends on

- students (s) (rating subjectively)
- teacher (d) – main interest
- department (dept) [[obfuscated]]
- "service" lecture or "own department student", (service: $0/1$).
- semester of student at time of rating (studage$\in \{2, 4, 6, 8\}$).
- how many semesters back was the lecture (lectage).

Main question: $\boxed{\text{Who's the best prof?}}$

Hence, for "political" reasons, want d as a **fixed** effect.

Model for ETH teacher ratings

- Want d ("teacher_ID", 1128 levels) as **fixed** effect.
- Further fixed effects studage (4 l.), lectage (6 l.), maybe service (2 l.), dept (15 l.).
- Use the new sparse.model.matrix() [1]:

```
> X <- sparse.model.matrix(~d + dept * service + studage +
+     lectage, data = d.eth)
> dim(X)

[1] 73421  1165
> object.size(X)/(ncol(X) * nrow(X) * 8)

[1] 0.021
> str(X)

Formal class 'dgCMatrix' [package "Matrix"] with 6 slots
  ..@ i       : int [1:867068] 0 1 2 3 4 5 6 7 8 9 ...
  ..@ p       : int [1:1166] 0 73421 73452 73485 73544 73574 73613 73641 73
  ..@ Dim     : int [1:2] 73421 1165
  ..@ Dimnames:List of 2
  .. ..$ : chr [1:73421] "1" "2" "3" "4" ...
  .. ..$ : chr [1:1165] "(Intercept)" "d6" "d7" "d8" ...
  ..@ x       : num [1:867068] 1 1 1 1 1 1 1 1 1 1 ...
```

Large fixed effect (ETH teacher)

- Now, in

  $$y = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b} + \boldsymbol{\epsilon}$$

  have $\boldsymbol{X}$ as $n \times 1165$, $\boldsymbol{Z}$ roughly $n \times 5000$, $n = 73'421$.

- using “developmental” `lmer2(..., sparseX = TRUE)` with *sparse* $\boldsymbol{X}$ (fixed effects) in addition to sparse $\boldsymbol{Z}$ (random effects)
  :

  ```
  > fm0 <- lmer2(y ~ d + dept * service + studage + lectage +
  +      (1 | s), data = d.eth, sparseX = TRUE)
  Error ... Cholmod error ’not positive definite’ at file:../Cholesky/..
  ```

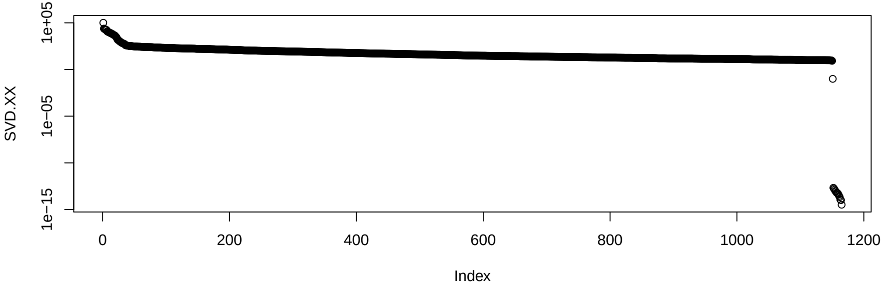
- indeed, fixed-effects $\boldsymbol{X}$ is rank-deficient, whereas making it random, *regularizes*

14 columns would have to be eliminated to render it non-singular:

```
> if (!exists("SVD.XX")) SVD.XX <- svd(crossprod(X), 0,
+     0)$d
> table(SVD.XX < 1e-05 * median(SVD.XX))

FALSE  TRUE
 1151    14

> plot(SVD.XX, log = "y")
```



- Finding: Do not use `'dept'` as that is entirely explained by `'d'`
- Then, the corresponding $\boldsymbol{X}$ has full rank ($p = 1137$)

```
> fm1 <- lmer2(y ~ d + service + studage + lectage + (1 |
+     s), data = d.eth, sparseX = TRUE)
## now call the minimizer -- on here one-dimensional:
str( optimize(fm1 @ setPars, c(0,5)) )
```

```
List of 2
 $ minimum  : num 0.278
 $ objective: num 235716
```

```
beta.fix <- get("beta", envir = lme4a::env(fm1)) # to keep ..
ls.str(lme4a::env(fm1))
```

```
beta :  Named num [1:1137] 3.807 -1.102 0.164 -1.272 -0.36 ...
contrasts :  NULL
control :  list()
data : 'data.frame': 73421 obs. of  7 variables:
 $ s      : Factor w/ 2972 levels "1","2","3","4",..: 1 1 1 1 2 2 3 3 3 3 .
 $ d      : Factor w/ 1128 levels "1","6","7","8",..: 525 560 832 1068 62 4
 $ studage: Ord.factor w/ 4 levels "2"<"4"<"6"<"8": 1 1 1 1 1 1 1 1 1 1 ...
 $ lectage: Ord.factor w/ 6 levels "1"<"2"<"3"<"4"<..: 2 1 2 2 1 1 1 1 1 1
 $ service: Factor w/ 2 levels "0","1": 1 2 1 2 1 1 2 1 1 1 ...
 $ dept   : Factor w/ 15 levels "5","12","6","11",..: 15 5 15 12 2 2 14 3 3
```

'd' as random effect: Regularization

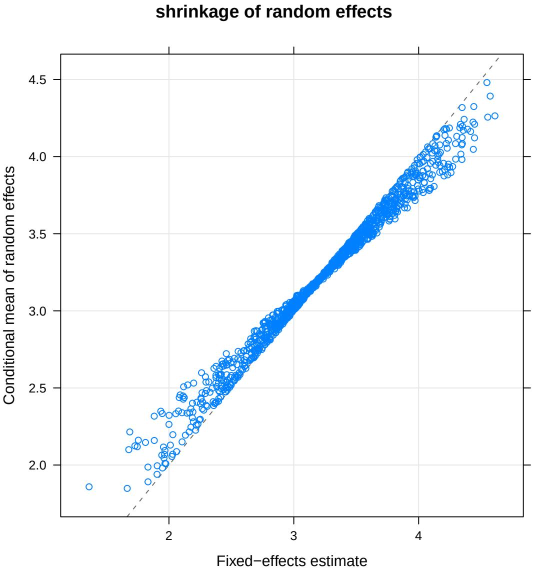
We rather prefer some bias in order to reduce variance:

make 'd' a random effect:

Advantage: Can easily use dept as a fixed effect:

```
> fm2 <- lmer(y ~ service + studage + lectage + (1 | d) +
+     (1 | s), data = d.eth)
> fm2D <- lmer(y ~ dept + service + studage + lectage +
+     (1 | d) + (1 | s), data = d.eth)
> beta.Fix <- beta.fix["(Intercept)"] + c(0, beta.fix[2:1128])
> b.random <- fixef(fm2)["(Intercept)"] + ranef(fm2)$d[,
+     1]
```

Shrinkage of the random effects relative to fixed effects


shrinkage of random effects
Conditional mean of random effects
Fixed−effects estimate