University of Zurich[UZH]

**EBPI** Epidemiology, Biostatistics and Prevention Institute

# Some things you can't read from a NEWS file

Torsten Hothorn

useR! 2024, Salzburg

# What to expect

- Statisticians try hard to prevent and eliminate bias.
- We are incarnations of Popper's "objective rationalist".
- For the next hour, I'll violate this principle by giving a personal, subjective, and most probably incorrect, account.

I picked the **mvtnorm** package I maintain as a Guinea pig.

# pkg:mvtnorm

Computes multivariate normal probabilities $\mathbb{P}(\mathbf{a} < \mathbf{Y} \leq \mathbf{b} \mid \mu, \mathbf{C})$ for $\mathbf{Y} \sim \mathbb{N}_J(\mu, \Sigma = \mathbf{C}\mathbf{C}^\top)$:

$$\sqrt{\frac{\det(\mathbf{C}^{-1})}{(2\pi)^{-J}}} \int_{\mathbf{a}}^{\mathbf{b}} \exp\left(-\frac{1}{2}(\mathbf{y}-\mu)^\top \mathbf{C}^{-\top}\mathbf{C}^{-1}(\mathbf{y}-\mu)\right) d\mathbf{y}$$

```
library("mvtnorm")
pmvnorm(lower = a, upper = b, mean, sigma, ...)
```

for unstructured covariance matrices $\Sigma$ in (relatively) large dimensions $J$.

# 1992

Alan Genz publishes paper in JCGS.

Core idea: compute Cholesky factor **C**, transform integral to $[0, 1]^{J-1}$, and use (quasi-) Monte-Carlo methods to evaluate numerically.

## Numerical Computation of Multivariate Normal Probabilities

ALAN GENZ*

The numerical computation of a multivariate normal probability is often a difficult problem. This article describes a transformation that simplifies the problem and places it into a form that allows efficient calculation using standard numerical multiple integration algorithms. Test results are presented that compare implementations of two algorithms that use the transformation with currently available software.

**Key Words:** Adaptive integration; Monte Carlo; Multivariate normal distribution.

# 1999

Alan Genz joins forces with Frank Bretz; extension to multivariate *t* distribution for multiple testing and simultaneous comparisons in the normal linear regression model.

## Comparison of Methods for the Computation of Multivariate *t* Probabilities

Alan GENZ and Frank BRETZ

This article compares methods for the numerical computation of multivariate $t$ probabilities for hyper-rectangular integration regions. Methods based on acceptance-rejection, spherical-radial transformations, and separation-of-variables transformations are considered. Tests using randomly chosen problems show that the most efficient numerical methods use a transformation developed by Genz for multivariate normal probabilities. These methods allow moderately accurate multivariate $t$ probabilities to be quickly computed for problems with as many as 20 variables. Methods for the noncentral multivariate $t$ distribution are also described.

**Key Words:** Multivariate $t$ distribution; Noncentral distribution; Numerical integration; Statistical computation.

# 2000

2000-11-14: **mvtnorm** 0.1-8 published on CRAN.

```
Package: mvtnorm
Title: Multivariate Normal and T Distribution
Version: 0.1-8
Author: Alan Genz <AlanGenz@wsu.edu>,
  Frank Bretz <bretz@ifgb.uni-hannover.de>,
  R port by Torsten Hothorn
  <Torsten.Hothorn@rzmail.uni-erlangen.de>
Description: computes the multivariate normal
  and t distribution
License: GPL
```

Why?

# Why mvtnorm?

Pro:

- I was a first year PhD student with a lot of time on my hands.
- My father was Franks PhD supervisor.
- I had been lobbying for R at family dinners for quite some time.
- Maybe I wanted to make a point here.

Con:

- Software, let alone a user interface, was not considered research output (JSS / R News / ... just gained speed).
- Open science was still behind the horizon.
- I was warned that such frills wouldn't land me a job at big pharma.

# Alan's prediction

```
Date: Thu, 16 Nov 2000 14:22:42 -0800
From: Alan C Genz
Subject: Re: mvtnorm

Torsten,

Thank you, I really appreciate the implementation
work that you have done.
[...]
This sounds like a good way to let more people know
about the software.

Best wishes,

Alan Genz
```

# Alan’s prediction

## Alan Genz

FOLGEN

Emeritus Professor of Mathematics, Washington State University
Bestätigte E-Mail-Adresse bei wsu.edu - Startseite

Numerical analysis

| TITEL | ZITIERT VON | JAHR |
|---|---|---|
| Numerical computation of multivariate normal probabilities<br>A Genz<br>Journal of computational and graphical statistics 1 (2), 141-149 | 1359 | 1992 |
| Computation of multivariate normal and t probabilities<br>A Genz, F Bretz<br>Springer Science & Business Media | 1112 | 2009 |
| mvtnorm: Multivariate Normal and t Distributions<br>A Genz, F Bretz, T Miwa, X Mi, F Leisch, F Scheipl, T Hothorn<br>R package version 0.9-2, URL http://CRAN. R-project. org/package= mvtnorm | 781 | 2008 |

# The times they were a-changin'

Pro:

- UZH Open Science policy:

  *UZH expects code and software created by UZH researchers to be shared with an adequate open source license.*

- Funding agencies consider software research output.
- JSS / R Journal / JOSS / ...
- R now does gets you a job at big pharma, at least in Basel.

Con:

- Software as research output is subject to bureaucratic scrutiny.
- Updating a package with 1008 reverse dependencies on CRAN sometimes causes headaches.

# Champagne!

```
Date: Tue, 21 May 2024 19:08:05 +0200
From: ligges@statistik.tu-dortmund.de
Cc: CRAN-submissions@R-project.org
Subject: [CRAN-pretest-publish] CRAN Submission

Dear maintainer,

thanks, package mvtnorm_1.2-5.tar.gz is on its
way to CRAN.

Best regards,
CRAN teams' auto-check service

Package check result: OK

No changes to worse in reverse depends.
```

# Package life

- Started as a simple R interface to Genz' and Bretz' monolithic FORTRAN code.
- Replaced some parts by R API (`runif()`, `pnorm()`).
- Other algorithms added (Miwa, TVPACK).
- Added features (random numbers, densities, improved quantiles, ...)
- R News article in 2001:

  *Using package **mvtnorm**, the efficient computation of multivariate normal or t probabilities is now available in R. We hope that this is helpful to users / programmers who deal with multiple testing problems.*

# Collaborators

- Alan Genz
- Frank Bretz
- Tetsuhisa Miwa
- Xuefei Mi
- Friedrich Leisch
- Fabian Scheipl
- Björn Bornkamp
- Martin Mächler

# 8639 days of package maintenance

[...] nothing but [...]

- blood,
- toil,
- tears, and
- sweat.

# 8639 days of package maintenance

[...] nothing but [...]

- bugs,
- unintended applications,
- feature requests, and lots of
- benefits.

# Bugs

```
Date: Sun, 23 Jul 2023 17:18:00 +0100
From: Prof Brian Ripley
Subject: CRAN package mvtnorm

See the logs at https://www.stats.ox.ac.uk/...

The warnings should be self-explanatory, but the
error requires the declaration

     DOUBLE PRECISION TVTMFN

ca tvpack.f: 62

Please correct before 2023-08-19 to safely retain
your package on CRAN (CRAN submissions are shut
until ca Aug 8).
```

# Unintended applications

```
Date: Thu, 15 Mar 2001 11:23:38 -0800
From: Timothy R. Johnson
Subject: Re: [R] tetrachoric correlations

I have a function that I wrote for R to compute
tetrachoric/polychoric correlations via maximum
likelihood.
[...]
negloglik <- function(theta) {
  [...]
  for(i in 1:n) {
   for(j in 1:m)
    prb[i,j] <- pmvnorm(c(0,0), sig, ...)
  }
  [...]
}
```

# Feature requests

```
Date: Fri, 08 Nov 2002 16:10:00 +1000
From: Fiona Evans
Subject: mvtnorm

Dear Torsten,

I've been experimenting with [...] and have found
the mvtnorm package to be very useful - thanks to
you and all involved in its development!

The amount of data in [...] is very large, it would
help if pmvnorm could read and process entire arrays
of data instead of just one data point at a time.
Is there any way the package can do this?

Regards, Fiona.
```

# Feature requests

```
Date: Thu, 21 Mar 2019 17:00:13 +0000
From: Devin Craig Francom
Subject: mvtnorm wishlist

Hi Torsten,

I was excited to find the pmvnorm function in the
mvtnorm package.
[...]
It would be nice if the user could pass the Cholesky
factor rather  than the covariance matrix as an
argument, [...]

All the best,
Devin
```

# Feature requests

The answer to these and similar requests was along the lines

> *Uhm, well, yes, this cannot be done easily.*

We will come back to these issues later.

# Depends: mvtnorm

- **multcomp**
- **coin**
- **party**
- **…**

# multcomp

$H_0 : \mathbf{K}\beta = 0$ based on $\mathbf{K}\hat{\beta}$ and $\mathbf{K}\hat{\Sigma}\mathbf{K}^\top$

```
library("multcomp")
summary(glht(<some model>(y ~ x, data),
             linfct = mcp(x = "<some contrast>")))
```

Essentially only `coef()` and `vcov()` are necessary, this works for many "linear" models, including GLM, Cox, mixed, robust etc.

**mvtnorm** does the leg work.

10.1002/bimj.200810425: 12'997 citations since 2008.

Simon Urbanek, Frank Bretz, TH, Achim Zeileis

# coin

```
library("coin")
independence_test(y ~ x, xtrafo = g, ytrafo = h)
```

computes min-$P$ based on Strasser and Weber's linear statistic

$$\sum_{i=1}^{N} g(\mathbf{x}_i) h(\mathbf{y}_i) \in \mathbb{R}^{p \times q}$$

Old and new test procedures, e.g. choosing $h$ as model residuals.

For $p \times q > 1$, **mvtnorm** does the leg work.

10.1198/000313006X118430 + 10.18637/jss.v028.i08:
2'389 citations since 2006.

Achim Zeileis, TH, Kurt Hornik

Henric Winell

# party

```
library("party")
varimp(cforest(y ~ ., data, ...))
```

Unbiased permutation variable importances for random forests.

10.1186/1471-2105-8-25: 3’793 citations since 2007.

TH, Carolin Strobl, Theresa Scharl, Bettina Grün, Fritz Leisch


Kid 2, Kid 1

# Lego systems

Identifying and implementing such frameworks is fun and helps statisticians to quickly put together models or inference procedures for designs where the stats cookbook lacks a tasty recipe.

Computational details are encapsulated and sometimes boil down to calling **mvtnorm**.

# The end of history?

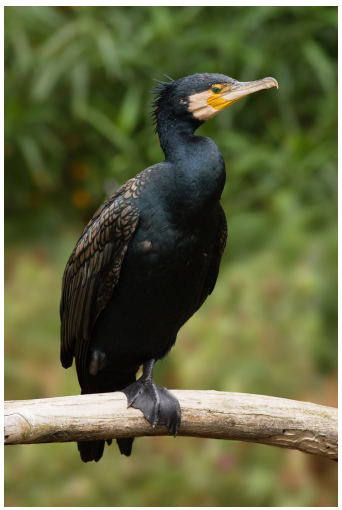
```
Changes in version 1.0-0 (2014-07-08)

After 14 years, we now feel safe enough to publish
mvtnorm 1.0-0. Many packages depend, import, or
suggest mvtnorm, so this version change also
indicates that the package is now stable and,
to a very large extent, the API is frozen. We
will of course continue to fix bugs or other
problems but new features are unlikely to go
into this package.
```

This didn’t stand the test of time, as Francis Fukuyama’s 1992 book.

# Aquatic birds

Great Crested Grebe

Goosander



Great Cormorant

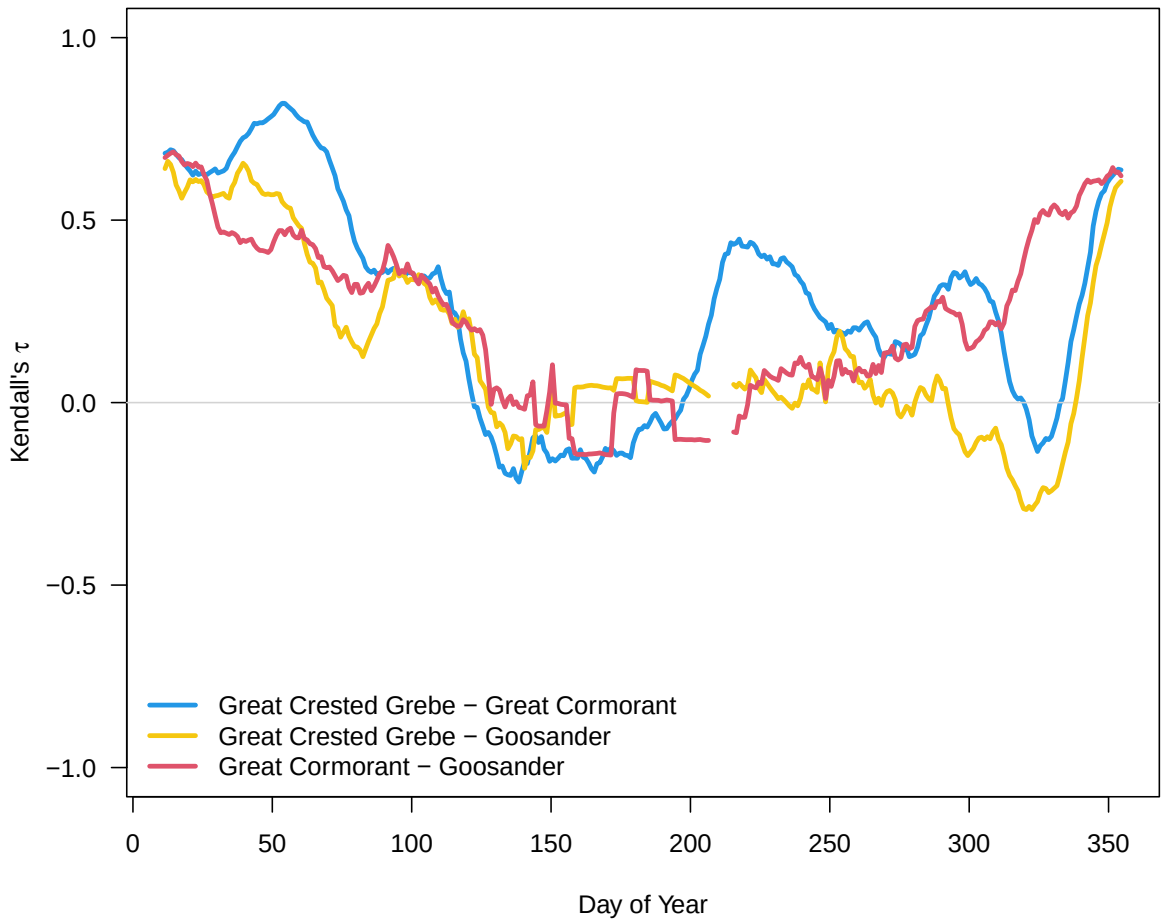
# Cooccurrences at Seehamer See

Daily data by Gerhard Kinshofer: 2002-05-01–2016-11-13.

Collaboration with Roland Brandl and Luisa Barbanti.

# Moving window Kendall's $\tau$

Time-dependent correlations?



# Model and likelihood for “polychoric” correlations

Latent Gaussian copula model for counts
$\mathbf{y} = (y_{\mathrm{GCG}}, y_{\mathrm{G}}, y_{\mathrm{GC}})^\top \in \mathbb{N}^3$ with time-varying correlations:

$$\mathbb{P}(\mathbf{Y} \leq \mathbf{y} \mid \mathrm{Year}, \mathrm{DoY}) = \Phi_3(\mathbf{h}(\mathbf{y} \mid \mathrm{Year}, \mathrm{DoY}) \mid \Sigma(\mathrm{DoY}))$$

Likelihood contribution:

$$\mathbb{P}\left( \begin{pmatrix} y_{\mathrm{GCG}} - 1 \\ y_{\mathrm{G}} - 1 \\ y_{\mathrm{GC}} - 1 \end{pmatrix} < \mathbf{Y} \leq \begin{pmatrix} y_{\mathrm{GCG}} \\ y_{\mathrm{G}} \\ y_{\mathrm{GC}} \end{pmatrix} \mid \mathrm{Year}, \mathrm{DoY} \right)$$

# Parameterisation

Marginal count models with transformation $\mathbf{h} = (h_{\text{GCG}}, h_{\text{G}}, h_{\text{GC}})^\top$

$$h_j(y_j \mid \text{Year}, \text{DoY}) = \frac{h_j(\lfloor y_j \rfloor)}{\sigma_j(\text{DoY})} - \mu_j(\text{DoY}) - \beta_{\text{Year}}$$

Gaussian copula:

$$\begin{aligned}
\mathbf{L}(\text{DoY}) &= \begin{pmatrix} 1 & 0 & 0 \\ \lambda_{21}(\text{DoY}) & 1 & 0 \\ \lambda_{31}(\text{DoY}) & \lambda_{32}(\text{DoY}) & 1 \end{pmatrix} \\
\mathbf{C}(\text{DoY}) &= \text{diag}(\mathbf{L}(\text{DoY})^{-1}\mathbf{L}(\text{DoY})^{-\top})^{-1/2}\mathbf{L}(\text{DoY})^{-1} \\
\Sigma(\text{DoY}) &= \mathbf{C}(\text{DoY})\mathbf{C}(\text{DoY})^\top
\end{aligned}$$

ensuring identifiability because $\text{diag}(\Sigma(\text{DoY})) \equiv 1$.

# Parameterisation

The shift ($\mu_j$), scale ($\sigma_j$), and inverse Cholesky ($\lambda_{jj}$) terms are modelled as a superposition of sinusoidal waves of different frequencies

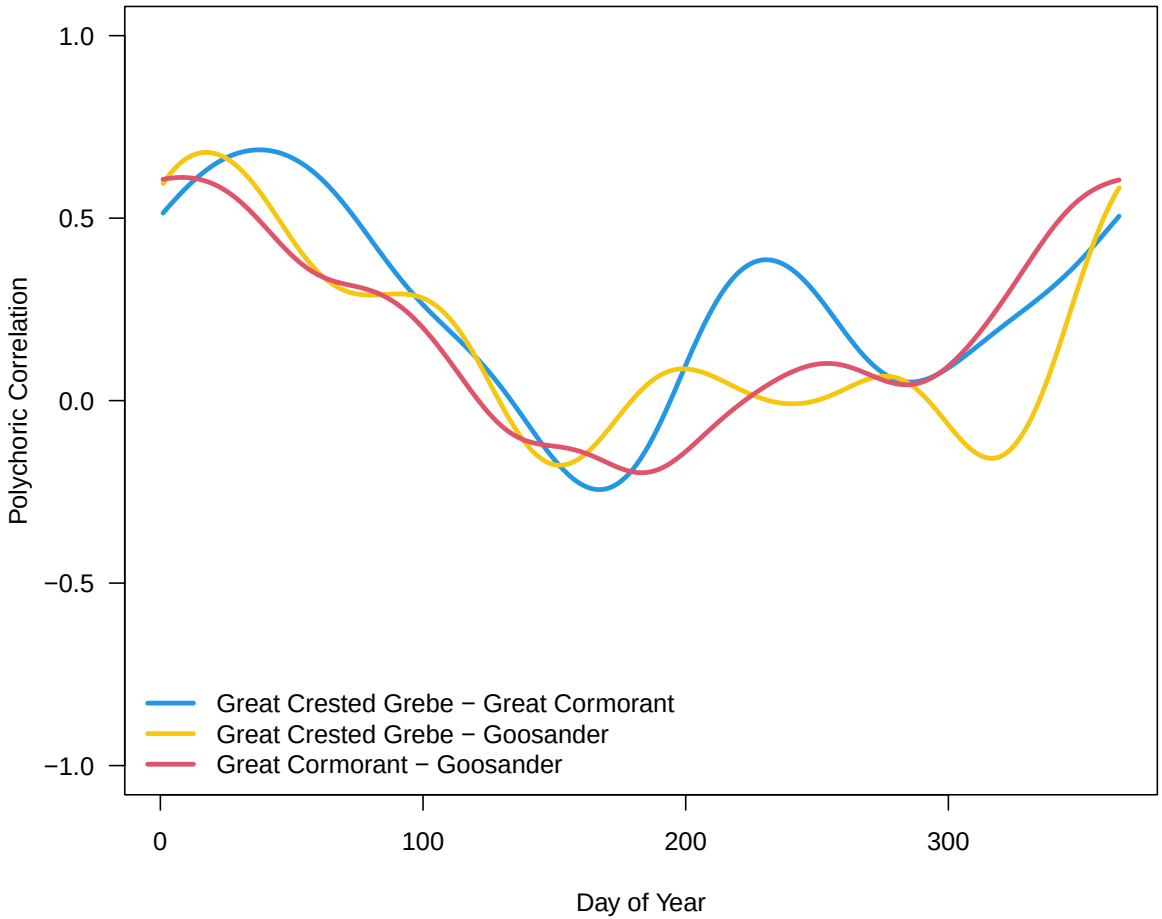
$$\sum_{s=1}^{4} \alpha_{1s} \sin\left(2 \times \pi \times s \times \frac{\text{DoY}}{365}\right) + \alpha_{2s} \cos\left(2 \times \pi \times s \times \frac{\text{DoY}}{365}\right)$$

such that the function values coincide on Dec 31 and Jan 1.

**h** is modelled by polynomials in Bernstein form of order 4.

The off-diagonal elements of $\Sigma(\text{DoY})$ are the polychoric correlations.

# Fitted model




LR test against constant $\lambda_{jj}$: $\chi^2_{24} = 511.65, P < .001$.

# mvtnorm 1.2-x

- $\mathbf{L}(\text{DoY}_i)$, $i = 1, \ldots, N = 4955$, is represented by an object of class "`ltMatrices`", a class for $N$ lower triangular matrices with methods (`solve()`, ...)
- Approximate log-likelihood *based on* **C**: `ldpmvnorm()`
- Score function (gradient of approximation): `sldpmvnorm()`
- Modular *vectorised* re-implementation (R and C) of Genz' 1992 procedure for log-likelihood and score functions, suitable for optimisation.
- Numerical integration with the help of **qrng** (Halton, Sobol, Korobov quasi-random sequences).
- All implementation details in

```
vignette("lmvnorm_src", package = "mvtnorm")
```

"`nuweb`" inspired by useR! 2016 keynote by Donald Knuth.

# tram: (Multivariate) most likely transformations

Multivariate transformation models of the form

$$\mathbb{P}(\mathbf{Y} \leq \mathbf{y} \mid \mathbf{x}) = \Phi_J\left(\mathbf{h}(\mathbf{y} \mid \mathbf{x}) \mid \Sigma(\mathbf{x})\right)$$

are rather general. Connections to DAGs, normalising flows and optimal transport exist.

```
library("tram")
m1 <- tram(y1 ~ x, ...)
m2 <- tram(y2 ~ x, ...)
...
mJ <- tram(yJ ~ x, ...)
mmlt(m1, m2, ..., mJ, formula = ~ x, data, ...)
```

Works for continuous, discrete, and censored data.

# Playing Lego again

Dependent censoring (10.1080/01621459.2022.2161387, Deresa & Van Keilegom) *almost* as simple as

```
mT <- Coxph(Surv(time, event) ~ x, data, ...)
mC <- Survreg(Surv(time, dropout) ~ x, data, ...)
mmlt(mT, mC, formula = ~ 1, data, ...)
```

More serious:

```
Compris(Surv(time, events) ~ x, data, ...)
```

Full likelihood, with plain ML standard errors.

# Playing Lego again

**bizicount: Bivariate Zero-Inflated Count Copula Regression Using R**

John M. Niehaus, Texas A&M University
Lin Zhu, Renmin University of China, Texas A&M University
Scott J. Cook, Texas A&M University
Mikyoung Jun, University of Houston

**OrdNor: An R Package for Concurrent Generation of Correlated Ordinal and Normal Data**

Anup Amatya, New Mexico State University
Hakan Demirtas, University of Illinois at Chicago

**Multivariate Normal Variance Mixtures in R: The R Package nvmix**

Erik Hintz, University of Waterloo
Marius Hofert, University of Waterloo
Christiane Lemieux, University of Waterloo

**mvord: An R Package for Fitting Multivariate Ordinal Regression Models**

Rainer Hirk, WU Wirtschaftsuniversität Wien
Kurt Hornik, WU Wirtschaftsuniversität Wien
Laura Vana, WU Wirtschaftsuniversität Wien

Covariate-dependent polychoric correlations, evaluation of approximations (pseudo or composite likelihoods), simultaneous inference for multiple marginal parameters, ...

# Lego bricks for statistics

- Objective rationalists need to be able to roam a rich model landscape representing hypotheses and potential alternatives.
- Identification of model frameworks, core tasks for their estimation and appropriate interfaces thereof has huge potential to facilitate high-level innovation (theoretically and in-silico).
- UseRs building such infrastructure make it more likely that software supports statisticians, rather than replacing them.
- We need a better understanding of how to communicate with fitted and unfitted models (`coef()`, `vcov()`, `predict()` is not enough).

# Its your turn now!

- Longterm package maintenance comes with a lot of benefits for developeRs and useRs.
- Longterm package maintenance comes with a lot of headaches for developeRs (and useRs).
- It is time to start your own longterm package project now, in preparation of your useR! 2048 keynote.

# Did R lobbying at family dinners work?

Family @ useR! 2006, Vienna

# Did R lobbying at family dinners work?

- My father became a vivid useR.
- He switched teaching biostatistics using R to his many students from the Global South.
- There is still much to be done bringing useRs from all over the world together.
- The R Project equips open societies (and those on the way) with open data analysis software, instrumental for defending liberal democracies against their enemies (Popper again, beefed-up a little).
- Engaging with geographically more distant useR communities brings us closer to the liberal utopian world the internet and free software movement promised in the 1990s, rather than the dystopian flavour we are witnessing thanks to social media, totalitarianism and big tech.