Getting started – software for statistical PhD students

This document lists a number of computer tasks which are useful for PhD students in Statistics, and itemizes software solutions for each of the tasks. We hope you will find the list useful, not only because it gives an impression of the range of possible software solutions for each task, but also because you may find it helpful to note the tasks themselves as useful ways to help you get the most out of your time as a PhD student – and beyond. The list was current as of *May 2010*; no doubt the range of possibilities will change as time passes so please feel free to let us know of software which you have found useful.

We have organized each of the task-lists in four columns, corresponding to Windows (XP, Vista, 7), Ubuntu (including Xubuntu, Kubuntu), Mac, and web-applications. These columns correspond to the basic choice of operating system which one makes when establishing a computer system. Much has been written about this choice, some of it helpful. Here are some basic pros and cons:

•**Windows. Pro:** often already installed. **Con:** not free, not open-source; applications often not well-integrated to command-line (a negative consideration for geeks); updating and adding new software is often clumsy.

•**Ubuntu** (we consider only one possible Linux implementation; there are many others and the considerations are essentially the same for most of them). **Pro:** free, open-source; excellent software repositories (Synaptic Package Manager) make it easy to update and to add most software; unix-based so good command-line integration (important for geeks). **Con:** some say Linux can lack final design polish; not many games (not important for geeks).

•**Mac. Pro:** excellent design; close linkage to hardware means it often "just works"; unix-based so good command-line integration (important for geeks). **Con:** not free, not open-source; close linkage to hardware means it is expensive.

•**Web. Pro:** doesn't depend on operating system; available *via* web connection whereever you are; someone else's responsibility to backup, maintain, and update. **Con:** is the privacy policy acceptable to you? How secure is the application in question? No access when offline?

Be aware of mix-and-match possibilities. Dual- or even triple-booting allows something of the best of all worlds. Linux systems can use wine to run Windows applications *within* a Linux environment (often a good solution for business software, can be rather less successful for games). The virtualbox application allows one to run complete operating systems within other operating systems. Windows users can install cygwin to access a huge number of unix programs, well-integrated into a unix command-line. But this is enough about operating systems; let us consider some of the fundamental tasks which you will need your computer to perform.

We have noted below where software is free, and where it is open-source. Free software has the obvious advantage of zero-cost; however a further advantage is that up-grades are usually free(!) and it can be installed on several machines without concern for licence conditions. Open-source software has some security advantages; in principle security holes can be publicly exposed and assessed, while the public availability of source-code gives some assurance that the software is really doing what it says it is doing. (However, both these advantages are often overstated.)

We have *not* given web links for these software packages. As ever, Google is your friend!

Email

Windows	Ubuntu	Mac	Web
Outlook Express	Evolution (Ubuntu: free, open source)	Mail	Gmail (free)
Outlook	Kmail (Kubuntu: free, open source)	Microsoft Entourage	Yahoo (free)
Thunderbird (free, open source) also Mutt (free, open source, for Geeks!)			

The first three columns list email clients, which can be used to read and / or download your email from your university or other email provider. It is often convenient to mix and match between the Web column and one of the other columns, as most email clients can be also configured to download email from a Web email application.

Calendar/Diary

Windows	Ubuntu	Mac	Web
Outlook	Evolution (Ubuntu: free, open source)	iCal	GCalendar (free)
	Korganizer (Kubuntu: free, open source)	Microsoft Entourage	
Rainlendar (free) or Mozilla Sunbird or Lightning (both free, open source)			

Do not underestimate the usefulness of running a computerized calendar / diary for forthcoming events! Good choices can email you reminders, display appointments for the month ahead (to check for deadline clashes), and serve as a record of what you did and when (vital at some later stage when you have to write activity reports). The habit of maintaining a calendar will serve you well throughout your professional life.

Computation

Windows	Ubuntu	Mac	Web
Mathematica			
Matlab			
Maple			
R (free, open source)			
Python (free, open source) also Octave (like Matleb) and Maxima (like Maple/Mathematica), both free and open source			

R is best of all of these for statistical purposes. Python is good for scripting and can interface with R. In general, computational speed-ups can often be obtained by investigating libraries / functions which work on matrices and vectors, rather than individual numerical quantities. Note that useful environments for R exist: in particular the celebrated cross-platform editor Emacs (free, open-source) can supply a very good R environment.

There are many many other possibilities too. Do not underestimate the merits of open-source software for scientific computation. Eventually you are likely to come to the point of needing to know *exactly* how a certain calculation is achieved, and then the ability to inspect the source-code can be invaluable.

Office software

Windows	Ubuntu	Mac	Web
Microsoft Office		Microsoft Office	Google Docs (free)
Openoffice (free, open source)			
		iWork	

If careful attention is paid to installing and working with the correct fonts, then Openoffice can be a very successful substitute for non-free alternatives.

Scientific document preparation

Windows	Ubuntu	Mac	Web
Miktex (free, open source)	TexLive (free, open source)	mactex (free, open source)	
Emacs with various addons (free, open source)			
Winedt (shareware)	Kile (free, open source)	TeXShop (free, open source)	
TeXnicCentre (free, open source)			
Texmaker (free, open source)			

Get used to using latex to prepare scientific documents. Any document containing any amount of mathematics will look far far better in latex! Using pdflatex and a latex package such as beamer, you can prepare pdf-based presentations which match and surpass Powerpoint in quality, and are easily accessible to others. Finally, a latex environment (as listed in last three rows for the three operating systems) pays dividends (a) when you can't remember the exact latex command, (b) when you want access to a template to start a new document, (c) when spell-checking, (d) at the compilation stage, when the environment can take care of the multiple latex runs needed to resolve labelling issues and so forth.

Bibliography

Windows	Ubuntu	Mac	Web
Bibtex (free, open source)			MR Lookup (free)
Mendeley (free)			Mendeley Web (free)
Jabref(free, open source)			Zotero (free)
		Papers BibDesk	

We should also note Papers (£25, mekentosj.com); some reckon this to be rather better at present

than Mendeley if you have a Mac. Speaking generally, get used to using bibtex to keep track of references to scientific papers. This uses a bibtex source file (which lists scientific papers in a curious format) to insert references and bibliography in a latex document. In practice you can build up a single bibtex source file containing all the articles you ever read, and use this again and again as you write various papers. The bibtex format is cumbersome; however you can typically download bibtex entries directly from the web (MR Lookup, Zotero, and other sources), and bibliography managers such as Mendeley, Jabref can be used to generate databases with graphical interfaces linking directly to PDFs of the relevant articles on your computer.

Note also the major online bibliographic databases that can export to BibTeX format, notably ISI Web of Knowledge (subscription service, most major universities subscribe; translation to BibTeX via a published Perl script, *isi2bibtex*) and Current Index to Statistics (subscription service, but with free access to records older than about 5 years), as well as Zetoc (free for scholars at a wide range of institutions) and Google Scholar (free).

Graphics

Windows	Ubuntu	Mac	Web
gimp (free, open source)			
dia (free, open source)			
xfig (free, open source)			
inkscape (free, open source)			

Gimp has capabilities approaching those of Photoshop, at least for those of us not involved in professional graphic design. Diagrams for papers can be generated quickly in dia or xfig, while inkscape produces vector graphics.

Back-up and synchronization

Windows	Ubuntu	Mac	Web
unison (free, open source)			dropbox (free for 2GB capacity)
duplicati (free, open source)	duplicity (free, open source)	Time Machine	

One day your computer's hard-disk is going to fail. What will you do then? Wise researchers ensure their work is backed-up onto an external disk, or *via* some web-service. The application unison will enable you to specifiy directories which you wish to be kept identical between computer and hard-disk, or indeed between two computers. Meanwhile duplicity and Time Machine can keep encrypted incremental backups, so that a given file can be accessed as it was last Tuesday evening. (NB: duplicati is a re-write of duplicity of which we have no direct experience).
One of us keeps 4 computers synchronized with unison, and uses duplicity to back-up incrementally (encrypted) to 2 separate hard-disks and a high-capacity memory stick, all based on habitually synchronizing at start and finish of each session, and backing-up once a week (prompted by emails from GCalendar). This habit has proved invaluable on three separate occasions in the last two years (each occasion resulting from a careless delete, rather than a disk crash).

Version control

Windows	Ubuntu	Mac	Web
rcs (free, open source)			
subversion (free, open source)			
git and Bazaar (free, open source)			

It is convenient to keep a record of the changes made to the paper on which you are currently working. In principle this can be recovered from a sequence of incremental back-ups; however it is simpler to use revision control. Every time one finishes a session of work on a document, one checks it in to the revision control system. Should one discover that the last week's work has involved accidentally deleting a magic paragraph, then this is easily recoverable by checking out an earlier version. Moreover one can use file comparison utilities to determine what parts have been changed recently. Once set up, these systems are easy to use, and invaluable when needed.

Source code editors (SCE), integrated development environments (IDE)

Windows	Ubuntu	Mac	Web
Emacs (free, open source)			
vi, Vim, gVim (free, open source)			
Notepad++ (free, open source)	gedit (free, open source)	Xcode (free)	
Microsoft Visual Studio	Kate (free, open source)		
	Kwrite (free, open source)		
	Kdevelop (free, open source)		
NetBeans (free, open source)			
Eclipse (free, open source)			

Most of you at some point in your careers will need to write some computer program. A source code editor (SCE) is an application that facilitates the process of editing source code for computer programs and in many cases also automates many steps in the programming process. Many SCEs are standalone applications (like the vi family, Notepad++, gedit, Kate, Kwrite, Kdevelop) while others are part of an Integrated Development Environment (IDE), which is an application that provides a comprehensive collection of tools (compilers, debuggers, source code editors, etc) that make one programmer's life easier. The table above provides a list of SCEs and IDEs that we have used in the past. One of us uses Emacs for all of his source code editing and scripting needs (R, C, C++, Python) and also for Latex. Actually, Emacs is so feature rich that it may be used for everything – email, web-browsing, file manager, ….

There is no such thing as the best SCE or IDE. We have met people who find the minimalistic interface of the vi family attractive, others who prefer using the feature-rich Emacs editor for as many things as possible and others that use a full-blown IDE for their programming needs. The general advice is to choose a SCE or IDE that fits your programming needs and try to understand why it does things the way it does. Then it will serve you well possibly for the rest of your career.

Further tasks

We have chosen not to cover a number of other tasks, for which we do not have much experience. For instance, the researcher of the future is likely to make much more use of blogging; will keep track of developments in the literature using an RSS newsreader, and will establish a professional presence using social websites. Be ready to make the most of such opportunities.