Cuadrados Mínimos Lineales

Assenza, Franco [1]

LU: 921/12

Maldonado, Kevin [2]

LU: 018/14

Umfurer, Alfredo [3]

LU: 538/12

Abstract

This is a short example to show the basics of using the ENDM style macro files. Ample examples of how files should look may be found among the published volumes of the series at the ENDM home page (`http://www.elsevier.com/locate/endm`)

Keywords: Please list keywords for your paper here, separated by commas.

1 Introducción

El análisis de la performance de una aerolínea es importante para distintos agentes: la aerolínea en sí, para analizar aspectos a mejorar; los clientes, para

[1] Email: `assenza.f@gmail.com`
[2] Email: `maldonadokevin11@gmail.com`
[3] Email: `aumfurer@gmail.com`

elegir la mejor opción a la hora de realizar un vuelo; la competencia, para tomar mejores decisiones de mercado; los gobiernos, para analizar la aerolínea de bandera y su competitividad, etcétera. El estudio de los retrasos de los vuelos es fundamental: vuelos retrasados pueden causar inconvenientes en la organización de los clientes, en sus combinaciones de vuelos. Retrasos excesivos pueden ser causas de resarcimientos económicos e influir negativamente en la imagen de la empresa.

Los algoritmos de aprendizaje automático nos permiten realizar predicciones de distintas variables en base a registros previos. En este caso, contamos con datos de registros de vuelos en Estados Unidos con información sobre la fecha, el origen y el destino, tiempos estimados y reales de despegue y aterrizaje, entre otros.

El análisis de series temporales tiene amplia difusión en la literatura (ver[10], por ej.), con aplicaciones importantes en muchas áreas del conocimiento. En este trabajo, vamos a enfocarnos en un modelo de regresión lineal para aprender sobre el cuerpo de datos y realizar un pronóstico a futuro.

1.1 Predicción

Para el modelo de predicción utilizamos regresión lineal[8], resuelto utilizando el método de Cuadrados mínimos lineales, con el algoritmo de[9] (pág 621) mediante la descomposición SVD.

2 Desarrollo

3 Implementación y ejecución

El método de cuadrados mínimos fue implementado en `C++` utilizando la librería de cálculo numérico `Eigen` para la descomposición SVD, los experimentos fueron realizados en `Python` utilizando las librerías de cálculo numérico `numpy` y `scipy`, y las librerías de visualización `matplotlib` y `seaborn`. Todo se ejecutó en una máquina con procesador `Intel(R) Core(TM) i7-6700K CPU @ 4.00GHz` y `32 GB` de memoria RAM.

3.1 Datos

Los datos pertenecen al *U.S. Departmen of Transportation* [4] , utilizados en una competencia de visualización de datos de la *American Statistical Association* [5] . Dado que los datos son susceptibles a problemas en Estados Unidos, en este trabajo estudiamos los datos a partir de 2002 (para evitar el ruido generado por el atentado a las torres gemelas) y antes de septiembre de 2008 (para no considerar los problemas de la recesión de 2008). Además, contamos con datos de los precios de las acciones de algunas aerolíneas, para buscar correlación entre estos y los datos del dataset de vuelos. Estos fueron extraídos de `Yahoo! Finance` [6] y se encuentran en `carrier_stock_prices/`.

3.2 Transformación

Al leer los datos, consideramos el retraso de un vuelo como: 0 si el vuelo está adelantado o no está atrasado, o el máximo entre el retraso de despegue y el retraso de aterrizaje (en minutos) con un valor máximo de 60 minutos. Estas decisiones fueron basadas en la mejor performance del algoritmo.

3.3 Extracción de features

La extracción de features se realizó de manera iterativa: se planteó una hipótesis (por ejemplo: "Los datos tienen una tendencia lineal creciente"), se ajustó una función acorde a la hipótesis (por ejemplo: una función lineal), se eliminó esa componente de los datos (por ejemplo: restarle a los datos el resultado de la función lineal) y se procedió a estudiar la siguiente hipótesis. Los features considerados fueron: tendencia lineal al crecimiento, periodicidad con distintas frecuencias, pico de retraso para días cercanos al 31/12, retraso medio de aerolíneas y aeropuertos (ver sección 3.4), día de la semana. .

3.4 Scores

Le asignamos a cada aerolínea un puntaje por año que es el estimado de retraso de la aerolínea ese año, y a cada aeropuerto dos puntajes por año, que son los estimados de retraso de ese aeropuerto como origen o como destino de vuelo. Los puntajes están entre 0 y 30 minutos como máximo.

---

[4] https://www.transtats.bts.gov/Fields.asp?Table_ID=236
[5] http://stat-computing.org/dataexpo/2009/
[6] https://finance.yahoo.com/

3.5 Evaluación

Para la evaluación del algoritmo se utilizó un conjunto de datos distinto al de entrenamiento, se convirtió a los tiempos de retrasos en booleanos: 0 si el retraso es menor que 15 minutos, 1 si el retraso es mayor o igual que 15 minutos, y sobre ese arreglo de booleanos se reportaron las métricas:

- Delay RMSE: RMSE de los tiempos de retraso de los vuelos, en minutos
- Delay NRMSE: RMSE de los tiempos de retraso de los vuelos, normalizado
- RMSE: RMSE del arreglo de booleanos
- Accuracy, precision, recall, balanced accuracy: del arreglo de booleanos

4 Resultados

4.1 Features

Realizamos un *scatter plot* de los datos, exponemos los resultados en la figura 1.

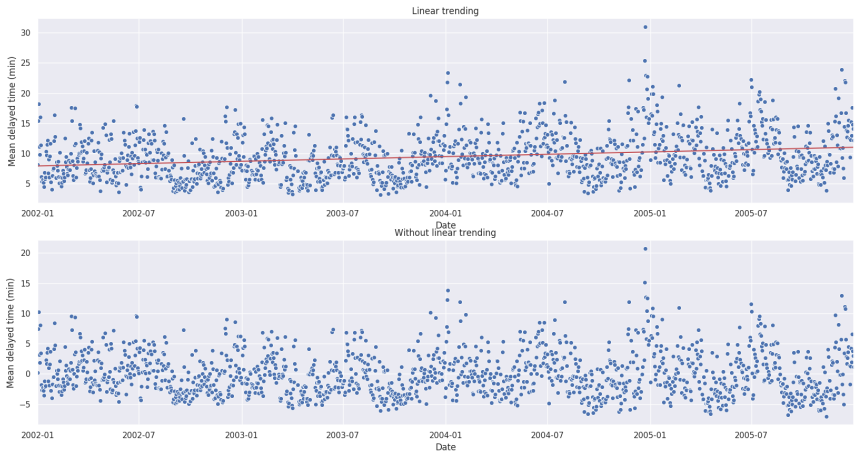


Fig. 1. *Detrending* de los datos

Realizamos un ajuste lineal de los datos, los que nos da un nuevo dataset con media cero (proceso conocido como *detrending*).

Luego, realizamos gráficos que nos muestren cómo se comportan los retrasos dentro de los meses de un año, los días de un mes y los días de una semana. Exponemos los datos en las figuras 2, 3 y 4.

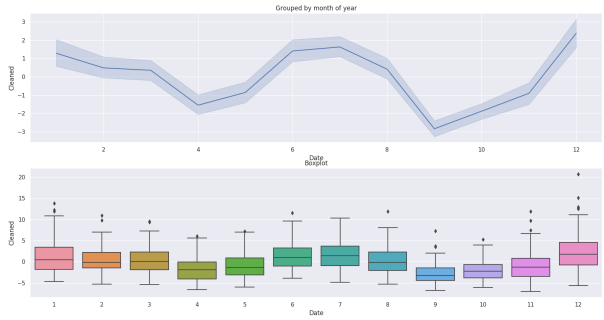


Fig. 2. Retraso medio por mes

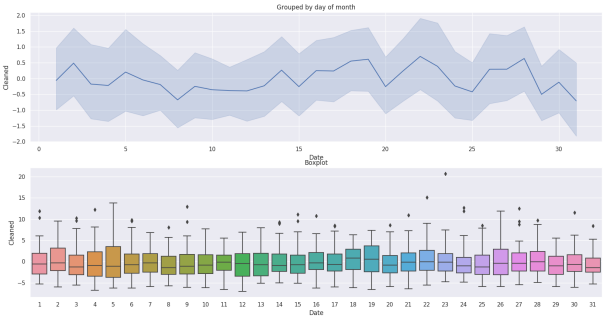


Fig. 3. Retraso medio por día del mes

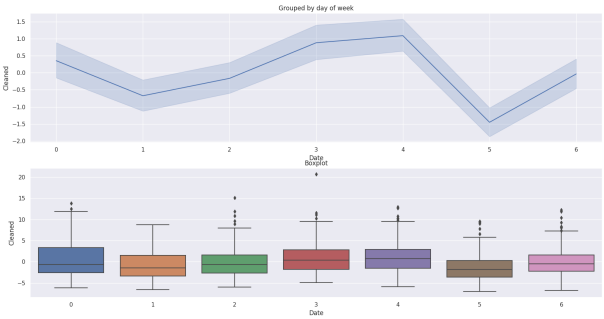


Fig. 4. Retraso medio por día de la semana

Las figuras muestran un comportamiento periódico de los retrasos. Es más claro el de la figura 2 que tiene elevaciones en los meses diciembre/enero y junio/julio, que es esperable pues son meses de vacaciones. También se ve

en la figura 4 un aumento en los días miércoles/jueves y domingo/lunes. El comportamiento parece ser menos predecible para la figura 3.

A modo de ejemplo, se muestran los más altos scores (explicados en la sección 3.4) del año 2008 en las figuras 5 y 6.

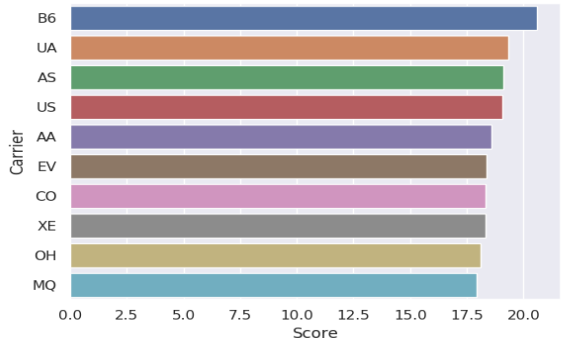


Fig. 5. Primeros puntajes de aerolíneas

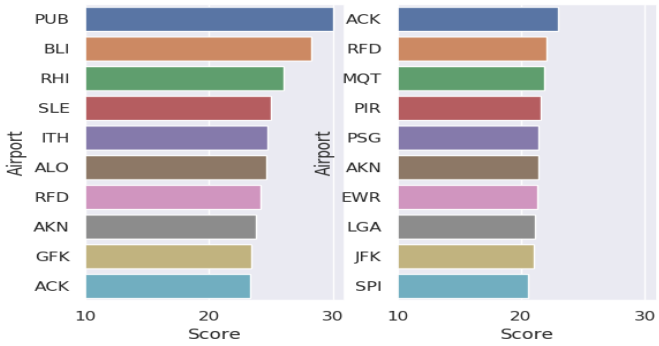


Fig. 6. Primeros puntajes de aeropuertos (origen y destino)

Sobre la base de este análisis se escogieron los features descritos en la sección 3.3. Se ajustó un modelo de cuadrados mínimos (sección 1.1). En la figura 7 se muestra el resultado del modelo con datos de entrenamiento entre 2002 y 2006 y datos de test a partir de 2006.

Nos planteamos la pregunta: ¿Aprende el algoritmo efectivamente de los datos? ¿Mejora la predicción si entrenamos con más datos? Para responder a esta pregunta, ejecutamos el modelo entrenando con una cantidad creciente de años, desde uno a todos menos uno, y testeamos contra los demás años. Exponemos los resultados de las métricas en la figura 8

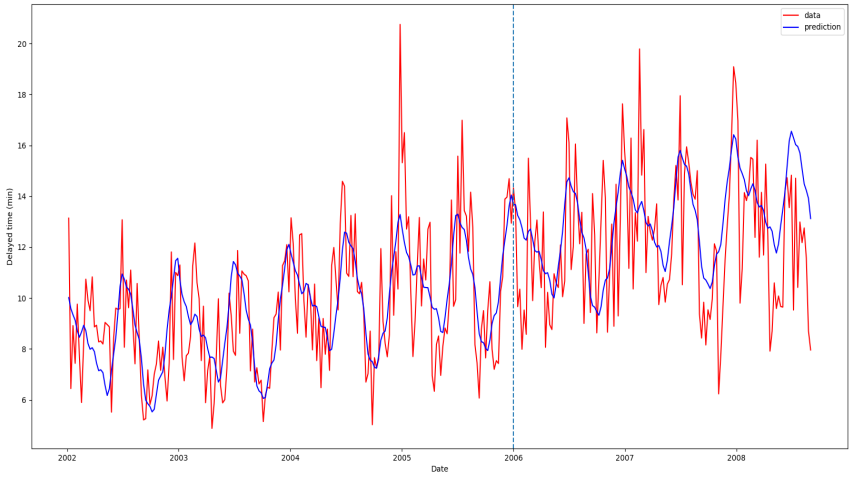


Fig. 7. Ejemplo de predicción. La línea punteada delimita el fin de los datos de entrenamiento y el comienzo de los datos de test

Training years	Delay RMSE	Delay NRMSE	RMSE	Accuracy	Precision	Recall	Balanced Accuracy
1	14.19	0.24	0.53	0.72	0.32	0	0.5
2	13.94	0.23	0.54	0.71	0.43	0	0.5
3	14.01	0.23	0.57	0.68	0.42	0.21	0.54
4	14.42	0.24	0.58	0.66	0.42	0.26	0.55
5	14.62	0.24	0.59	0.65	0.42	0.27	0.55
6	14.81	0.27	0.6	0.64	0.40	0.34	0.55

Fig. 8. Métricas para distintos datos de entrenamiento

Los resultados de las ejecuciones para distintos conjuntos de entrenamiento son paradójicos, ya que algunas métricas mejoran (como Recall y Balanced Accuracy), pero otras empeoran (como Delay NRMSE). Nuestra hipótesis es que se puede deber a que el modelo con más datos de entrenamiento es más susceptible a *overfitting*, o que al entrenar con conjunto de datos más grande se toma en consideración datos viejos que no son tan influyentes en los datos más nuevos.

En la figura 9 se grafican la cantidad de vuelos y el precio de las acciones de algunas aerolíneas. En ella se puede ver algunas cosas notables, como por ejemplo el gran desplome de los precios de las acciones con la crisis de 2008,

o la aparición importante de la aerolínea `AA`, y su caída en el 2008 aún más grande que la de las otras aerolíneas.

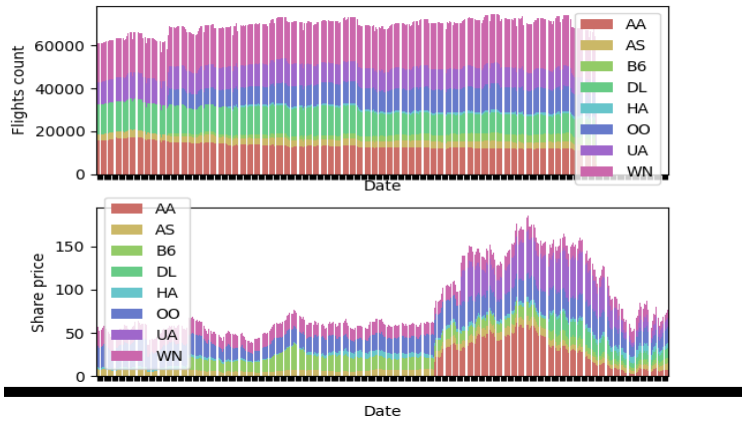


Fig. 9. Cantidad de vuelos y precio de las acciones por aerolínea

Una pregunta que nos planteamos es: ¿Están correlacionadas ambas variables? Para estudiarlo, consideramos, para cada aerolínea, qué porcentaje de vuelos le corresponda a esa aerolínea, y qué porcentaje de precio de la acción le corresponde (sobre el conjunto de aerolíneas para las que tenemos datos de precios de acciones), y estudiamos la correlación (vía coeficiente de Pearson) de ambas variables a lo largo del tiempo. Exponemos los resultados del coeficiente de Pearson en la figura 10. De la figura se desprende que no hay una correlación evidente entre estas variables, ya que los valores del coeficiente de Pearson son variables entre 0.74 y -0.94.

Carrier	Pearson
AA	0.74
AS	0.33
B6	-0.94
DL	0.53
HA	0.05
OO	-0.6
UA	0.29
WN	-0.62

Fig. 10. Coeficiente de Pearson entre el porcentaje de vuelos y el porcentaje de precio de las acciones para cada aerolínea

5 Frontmatter

The biggest difference between a "usual" LaTeX style such as `article.sty` and the ENDM package is that the ENDM macro package requires the title, author's name or names, abstract, keywords and "thanks" all to be included within the `frontmatter` environment. At the beginning of the source file for this paper, you'll notice this. Also, you'll notice that the usual `\maketitle` is absent; it no longer is needed. The ENDM style package automatically generates the title, author's name and address, and related material at the beginning of the paper. Note also that hyperref has been disabled in this part of the endm.cls file, so references to footnotes aren't linked to the appropriate footnotes or addresses. This is an old problem with LaTeX, involving the fact that the references within the frontmatter aren't passed cleanly to the linking software.

For those who have used the ENDM package before, the one new thing to note is the inclusion of *Keywords* which are now required by Elsevier.

The ENDM macro package provides two alternatives to listing authors names and addresses. These are described in detail in the file `instraut.pdf`. Basically, listing each author and his or her address in turn, is the simplest method. But, if there are several authors and two or more share the same address (but not all authors are at this address), then the method of listing authors first, and then the addresses, and of referencing addresses to authors should be used.

Furthermore, note that an acknowledgment of support (the contents of `\thanks`) should be done by a separate listing of `\thanks[NSF]{To the NSF}` with the optional argument – `[NSF]` – being used for `\thanksref` which is attached to those authors acknowledging such support. It is important that the `\thanks` not be included within the scope of `\author{}` or of `\title{}`, but it must be within the scope of the environment `frontmatter`.

More details about added terms such as `\collab` can be found in the file `instraut.pdf` if they are needed.

6 Sectioning and Environments

Since ENDM is published through the auspices of Elsevier B. V., their style files were used to create the ENDM macro package. Below is a proof which shows that this package is not much different to most others:

Definition 6.1 A file is *derived* from another file if it was obtained by making

only a few modifications to the original file.

Theorem 6.2 *The file* `endm.cls` *is derived from* `elsart.sty`.

Proof. This is clear from the similarity of the output to the output from the standard Elsevier style files. □

If one wants to start a proof with a descriptive word, such as "sketch", then one can use the `\begin{proof*}...\end{proof*}` environment, as in

Proof (Sketch) This can be derived from simple observations. □

The main difference between the file `endm.cls` and the `elsart.cls` file used for other Elsevier journals is the more precise format we use. Elsevier's generic style files are meant for preliminary editing and more precise formatting is imposed using a macro file designed for the specific Elsevier journal in which the paper will eventually appear. The `endm.cls` and `endmmacro.sty` files format papers uniformly so that they all are easily recognizable as belonging to the series *Electronic Notes in Discrete Mathematics*.

All of the usual features of LaTeX are available with these style files. It is only the formatting that has been rigorously defined. One can use the sectioning commands `\section,\subsection, \paragraph` and `\subparagraph`. The numbering scheme used is one under which Theorem 1.2.3 is the third numbered item in the second subsection of the first section of the paper. In order to facilitate cross-references, all of the named environments given below are numbered and all use the same numbering scheme.

The file `endmmacro.sty` contains additional information that is needed to typeset a paper. It also has the definitions of the $\mathcal{AMS}$ `euler` and `blackboard bold` fonts builtin. If you want to use symbols for the natural numbers, the reals, etc., then we prefer that you use the blackboard bold fonts, and not plain bold fonts. This is accomplished by using the `\mathbb` font, as in $\mathbb{N}$ or $\mathbb{R}$.

The names of theorem-like environments are provided in `endmmacro.sty`. With the exception of the environment "Algorithm", the names of all these are the full name rather than a shortened version. The environments provided and their names are as follows:

- `\begin{theorem} ... \end{theorem}` for Theorems,
- `\begin{lemma} ... \end{lemma}` for Lemmas,
- `\begin{corollary} ... \end{corollary}` for Corollaries,
- `\begin{proposition} ... \end{proposition}` for Propositions,

- `\begin{criterion} ... \end{criterion}` for Criteria,
- `\begin{alg} ... \end{alg}` for Algorithms,
- `\begin{definition} ... \end{definition}` for Definitions,
- `\begin{conjecture} ... \end{conjecture}` for Conjectures,
- `\begin{example} ... \end{example}` for Examples,
- `\begin{problem} ... \end{problem}` for Problems,
- `\begin{remark} ... \end{remark}` for Remarks,
- `\begin{note} ... \end{note}` for Notes,
- `\begin{claim} ... \end{claim}` for Claims,
- `\begin{summary} ... \end{summary}` for Summary,
- `\begin{case} ... \end{case}` for Cases, and
- `\begin{ack} ... \end{ack}` for Acknowledgements.

For example,

Algorithm 1 *Step 1: Write the paper*
Step 2: Format it with the ENDM macro package
Step 3: Ship the whole thing to the Guest Editors

7 References and Cross-references

All the cross-referencing facilities of LaTeX are supported, so one can use `\ref{}` and `\cite{}` for cross-references within the paper and for references to bibliographic items. As is done in this note, the *References* section can be composed with `\begin{thebibliography}...\end{thebibliography}`. Alternatively, BibTeX can be used to compile the bibliography. Whichever one is used, the references are to be numbered consecutively, rather than by author-defined acronyms. Of course you can use your own acronyms for easy reference to each of the items in the bibliography, as has been done with the listing for this short note.

Note that the references should *not* be started with a new `\section` command.

The package `hyperref` is automatically loaded by endm.cls and this makes all the cross-references within the document "active" when the pdf file of the paper is viewed with Adobe's Acrobat© Reader. The format for including a link is simple: simply insert `\href{URL} {text}` where *URL* is the URL

to which you want the link to point, and *text* is the text you want to be highlighted and which will bring up the desired web page when clicked upon.

7.1 Particulars about `.pdf` files

We now require that `.pdf` files be provided for publication online. A `.pdf` file is viewable by Adobe's Acrobat© Reader, which can be configured to load automatically within a browser. Viewing a properly formatted `.pdf` file with Acrobat© allows the cross-references and links to URLs to be active. In fact, Elsevier utilizes `.pdf` files in order to take better advantage of the web's capabilities.

One point that needs to be emphasized is that you should use Type 1 fonts when you typeset your LaTeX source file. These fonts are scalable, meaning that they carry information that allows the device viewing the final output to scale the fonts to suit the viewer being used (from an onscreen viewer such as Adobe's Acrobat© Reader to printing the file on a printer). You can tell if you have used the right fonts by viewing the final output on your machine. If the font looks grainy, then you have not used a Type 1 font. Type 1 fonts can be located at the CTAN archive at http://www.ctan.org. They are public domain fonts and do not cost anything when you add them to your system.

Assuming you have Type 1 fonts available, there are several methods for producing `.pdf` files.

Using `dvips` and `ps2pdf`

We list this option first since it appears to be the most reliable and the easiest to use, especially if you include embedded PostScript graphics (`.eps` files) in your source file. Simply run LaTeX2e on your source file, apply `dvips` to produce a PostScript file and then apply `ps2pdf` to obtain a `.pdf` file.

The `DVIPDFM` utility

Another easy method for producing acceptable `.pdf` files is via the utility `dvipdfm`. This utility is included in distributions of MikTeX, which runs on Windows machines, but it probably needs to be added to your teTeX distribution, if you are running LaTeX on a UNIX machine. The utility and precise information about installing it on your system can be found at the web page http://gaspra.kettering.edu/dvipdfm/. In essence, this utility converts a `.dvi` file into a `.pdf` file. So, one can first prepare the `.dvi` file using

LaTeX, and then apply the utility `dvipdfm` to produce the needed `.pdf` file.[7] This utility makes the inclusion of graphics particularly simple. Those that are included in the LaTeX source file are simply converted to the `.pdf` format. As we note below, things are not so simple with the second alternative, which is to use pdfLaTeX.

pdfLaTeX

An alternative to the first possibilities to produce `.pdf` files is to process the source file with pdfLaTeX. This format is available from the standard CTAN sites `http://www.ctan.org`. It appears that pdfLaTeX and `hyperref` have some problems when used together. It is necessary to use pdfLaTeX version 14d or later in order to minimize these issues. If your system has an earlier version (most teTeX distributions have version 13d), then you can update your system by retrieving the latest version of pdfLaTeX from `ftp://ftp.cstug.cz/pub/tex/local/cstug/thanh/pdftex/`. Even if the recent versions are used, pdfLaTeX has the same dealing with references embedded with the `frontmatter` section described above for LaTeX.

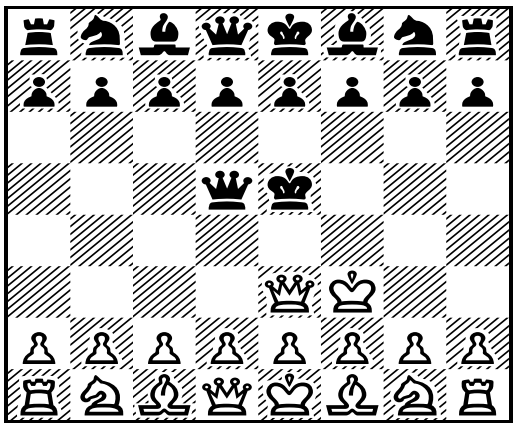
But there is one aspect of pdfLaTeX that creates problems. Many authors include EPS[8] files within their papers. While this is fairly straightforward with LaTeX, there are a couple of points to note when attempting this with pdfLaTeX.

To include a PostScript image in a `.pdf` file produced with pdfLaTeX, you first have to convert the image to a `.pdf` file. The conversion can be accomplished most easily using Ghostscript; you can simply view the file in Ghostview and then print the image to a `.pdf` file using the `pdfwriter` option within Ghostview. The result for a standard chess board that is part of the Ghostview distribution is the following image:

---

[7] *Beware*! The utility `dvipdf` does *not* produce acceptable `.pdf` files, and should not be used. Only `dvipdfm` should be used to produce `.pdf` files.

[8] EPS stands for *embedded PostScript*, which affords a mechanism for including pre-prepared PostScript files within a LaTeX document.



Below is a copy of a color image. While pdfLaTeX can handle image files in other formats, LaTeX can only handle `.eps` images reliably.

Using ENDM Macros with Mac OS X

Clearly, if your file does not require `.eps` or other PostScript files, then you can create the required `.pdf` file using any of the standard TEX implementations for the Macintosh. If you do need to include PostScript files and if you are using TEXShop, then you can specify to use dvips and Ghostview in processing your file, and then you can apply `ps2pdf` to create the needed `.pdf` file. Alternatively, the Mac OS X operating system is based on UNIX, so it supports the use of teTEX as described above.

8 Summary and Remarks

The ENDM macro package is relatively easy to use and provides a uniform layout for all the papers that appear in ENDM.

Assigning Volume Numbers

An additional point worth mentioning is that ENDM has moved to *ScienceDirect*, Elsevier's main platform for publishing electronic series. Because *ScienceDirect* cannot easily accommodate changes to published material, the *Proceedings* must be entirely ready before they can be published. Volume numbers will therefore not be assigned for the *Proceedings* until the final versions of all papers are in.

Copyright Transfer Forms

Due to the move to *ScienceDirect*, the corresponding author of each paper published in ENDM must submit a signed Copyright Transfer Form to Elsevier in order for their paper to be published. A copy of this form will be sent to each author. Note that the publication of an abstract or extended abstract in ENDM will not restrict the author(s) from publishing a full-length article on the same topic and with the same title in another journal (possibly with another publisher). Details about the copyright agreement specifying the exact rights of the authors and the rights of Elsevier are available at Elsevier's Author Gateway.

9 Bibliographical references

ENDM employs the `plain` style of bibliographic references in which references are numbered sequentially and listed in alphabetical order according to the first author's last name. Please utilize this style. We have a BibTEX style file,

for those who wish to use it. It is the file `endm.bst` which is included in this package. The basic rules we have employed are the following:

- Authors' names should be listed in alphabetical order, with the first author's last name listed first followed by initials or first name, and with the other authors' names listed as *first name, last name*.
- Titles of articles in journals should be in *emphasized* font.
- Titles of books, monographs, etc. should be in quotations.
- Journal names should be in plain roman type.
- Journal volume numbers should be in boldface, immediately followed by the year of publication enclosed in parentheses in roman type.
- References to URLs on the net should be "active" and the URL itself should be in `typewriter` font.
- Articles should include page numbers.

The criteria are illustrated by the examples below.

References

[1] Civin, P., and B. Yood, *Involutions on Banach algebras*, Pacific J. Math. **9** (1959), 415–436.

[2] Clifford, A. H., and G. B. Preston, "The Algebraic Theory of Semigroups," Math. Surveys **7**, Amer. Math. Soc., Providence, R.I., 1961.

[3] Freyd, Peter, Peter O'Hearn, John Power, Robert Tennent and Makoto Takeyama, *Bireflectivity*, Electronic Notes in Theoretical Computer Science **1** (1995), URL: http://www.elsevier.com/locate/entcs/volume1.html.

[4] Easdown, D., and W. D. Munn, *Trace functions on inverse semigroup algebras*, U. of Glasgow, Dept. of Math., preprint 93/52.

[5] Roscoe, A. W., "The Theory and Practice of Concurrency," Prentice Hall Series in Computer Science, Prentice Hall Publishers, London, New York (1198), 565pp. With associated web site http://www.comlab.ox.ac.uk/oucl/publications/books/concurrency/.

[6] Shehadah, A. A., "Embedding theorems for semigroups with involution, " Ph.D. thesis, Purdue University, Indiana, 1982.

[7] Weyl, H., "The Classical Groups," 2nd Ed., Princeton U. Press, Princeton, N.J., 1946.

[8] Hastie, T., "The Elements of Statistical Learning: Data Mining, Inference, and Prediction," 2nd Ed., Springer, 2019.

[9] Burden, R., and Faires, J., "Numerical Analysis," 9th Ed., Brooks/Cole, 2011.

[10] Shumway, R., and Stoffer, D., "Time Series Analysis and Its Applications," 3rd Ed., Springer, 2010.