Random Effect Models for Parton Distribution Functions?

Steffen Lauritzen, University of Oxford

Banff workshop

July 21, 2010

Experiments $i = 1, \ldots, m$; data in each experiment $j = 1, \ldots, n_i$;
Standard model

$$\chi^2 = \gamma \sum_{ij} \frac{\{\text{data}_{ij} - \text{theory}(\theta)_{ij}\}^2}{\text{error}_{ij}^2}$$

with θ being a vector of parameters and $\gamma = \sigma^{-2}$ potentially a scale factor for the error, acknowledging that the error model might be wrong by such a factor.

The standard model ignores that in every experiment the theory does not quite fit, so that each experiment should have its own parameter vector θ and it therefore grossly underestimates the error of predicting the results in a future experiment.

A *random effects model* formalises this by letting θ_i, the parameters in experiment i, be different but taken from a (population) distribution of parameter values, for example by assuming a joint distribution with log density proportional to

$$\tilde{\chi}^2 = \sum_{ij} \gamma \frac{\{\text{data}_{ij} - \text{theory}(\theta_i)_{ij}\}^2}{\text{error}_{ij}^2} + \lambda(\theta_i - \theta)^\top H(\theta_i - \theta),$$

i.e. it says that $\theta_i \sim \mathcal{N}\{\theta, (\lambda H)^{-1}\}$.

So, the formal parameters of this model are θ, possibly γ and λ, and even possibly H. The second term in the modified χ^2 represents and error type which, following Thiele (1880), could be termed 'quasi-systematic', see also Lauritzen (1981, 2002).

For simplicity consider the case where $\sigma^2 = 1$ is known and where we choose H to be the Hessian matrix of the first χ^2. This leaves λ as the single unknown parameter in the model. This may be slightly ad hoc as H cannot then be specified independently of the measurements. Should like to explore this choice further.

λ can then for example be estimated by maximum likelihood by maximizing

$$L(\lambda) = \int \exp\{-\tilde{\chi}^2\} \prod_{i=1}^{m} d\theta_i$$

which is a high-dimensional integral.

$L(\lambda)$ in general be maximised by using the EM algorithm, calculating

$$q(\lambda) = \mathbb{E} \log L(\lambda) = -\int \tilde{\chi}^2 \prod_{i=1}^{m} d\theta_i$$

by Monte–Carlo integration, then maximizing $q(\lambda)$ and iterating.

Full Bayesian analysis by MCMC is also possible and possibly preferable.

This type of analysis is known under many different names, each having its own little twist or focus of interest. Common names for a Google scholar search would be mixed models, mixed effect models, empirical Bayes, variance component models, multi-level models, hierarchical Bayes models, etc...

The original sources for empirical Bayes methods are Robbins (1956, 1964); an excellent overview and explanation of the merits of the methodology is given in Efron (2003); see for example also Gelman et al. (2004, chap. 5,chap. 15) and/or Carlin and Louis (2009, chap. 5) (Chapter 3 in second edition).

One interpretation of the methodology is that the second term in $\tilde{\chi}^2$ represents a Gaussian prior distribution of the parameters of each individual experiment, with covariance matrix $(\lambda H)^{-1}$.

It may be more adequate, although computationally typically more involved, to use a prior distribution with heavier tails, such as, for example a multivariate t-idstribution, or a distribution with density proportional to

$$\exp -\lambda\sqrt{(\theta_i - \theta)^\top(\theta_i - \theta)}.$$

Carlin, B. P. and Louis, T. A.: 2009, *Bayesian Methods for Data Analysis*, 3 edn, CRC Press, Boca Raton.

Efron, B.: 2003, Empirical Bayes and microarrays, *The Annals of Statistics* **31**, 366–378.

Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B.: 2004, *Bayesian Data Analysis*, 2 edn, Chapman and Hall/CRC Press, Boca Raton.

Lauritzen, S. L.: 1981, Time series analysis in 1880: A discussion of contributions made by T. N. Thiele, *International Statistical Review* **49**, 319–331.

Lauritzen, S. L.: 2002, *Thiele: Pioneer in Statistics*, Oxford University Press, Oxford.

Robbins, H.: 1956, An empirical Bayes approach to statistics, *Proceedings of the Third Berkeley Symposium in Mathematical Statistics and Probability*, Vol. 1, University of California Press, Berkeley, pp. 157–163.

Robbins, H.: 1964, The empirical Bayes approach to statistical decision problems, *The Annals of Mathematical Statistics* **35**, 1–20.

Thiele, T. N.: 1880, Om Anvendelse af mindste Kvadraters Methode i nogle Tilfælde, hvor en Komplikation af visse Slags uensartede tilfældige Fejlkilder giver Fejlene en 'systematisk' Karakter, *Det kongelige danske Videnskabernes Selskabs Skrifter, 5. Række, naturvidenskabelig og mathematisk Afdeling* **12**, 381–408. French version: *Sur la compensation de quelques erreurs quasi-systématiques par la méthode des moindres carrés.* C. A. Reitzel, København, 1880.