Confidence envelopes and surfaces for multiple regression fits in analytical chemistry

Ian Kelly and Paul J. Ogren* **Earlham College, Richmond, IN**

Curve or surface fitting to functions such as (1) or (2) is widely used in data analysis. The computations needed to construct best fit curves and surfaces can be easily extended to calculate confidence envelopes and surfaces for any functions within these classes.

- Confidence envelopes bound all of the alternative curves which might fit a data set within a desired level of confidence.
- Confidence envelopes provide a quick visualization of regions where a calibration is well defined or poorly defined.
- Confidence envelopes can be used to estimate detection limits and upper ranges of validity in a calibration.
- Confidence envelopes are most commonly used in straight line fits. This poster shows how to construct them for many other cases.

$$f(x_i) = a_1 X_1(x_i) + a_2 X_2(x_i) + a_3 X_3(x_i) + \ldots \qquad (1)$$

$$z(x_i, y_i) = a_1 Q_1(x_i, y_i) + a_2 Q_2(x_i, y_i) + a_3 X_3(x_i, y_i) + \ldots \qquad (2)$$

Changes in (1) due to changes in the parameters away from their best fit values are given by:

$$g(x_i) = \delta f(x_i) = \delta a_1 X_1(x_i) + \delta a_2 X_2(x_i) + \delta a_3 X_3(x_i) + \ldots \qquad (3)$$

It can be shown that the confidence envelope equations are then given by:

$$\begin{aligned} f_{env}(x_i) &= f(x_i) \pm g(x_i)_{\mathrm{max,min}} \\ &= f(x_i) \pm \left\{ F(\mathbf{L})^{\mathbf{T}} \left[(\mathbf{X})^{\mathbf{T}} (\mathbf{X}) \right]^{-1} (\mathbf{L}) \right\}^{\frac{1}{2}} \qquad (4) \\ &= f(x_i) \pm \left\{ F(\mathbf{L})^{\mathbf{T}} [\varepsilon] (\mathbf{L}) \right\}^{\frac{1}{2}} \end{aligned}$$

where $F \equiv pF^{1-\alpha}_{p,n-p}$ $(\approx \Delta\chi^2$ for large data sets).

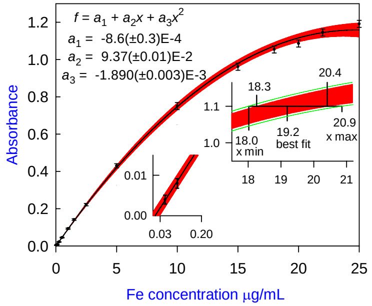
The vector **(L)** and the matrix **(X)** are given by:

$$(\mathbf{L}) = \begin{pmatrix} X_1(i) \\ X_2(i) \\ X_3(i) \\ \ldots \end{pmatrix} \qquad (\mathbf{X}) = \begin{pmatrix} X_1(1)/\sigma_1 & X_2(1)/\sigma_1 & X_3(1)/\sigma_1 & \ldots \\ X_1(2)/\sigma_2 & X_2(2)/\sigma_2 & X_3(2)/\sigma_2 & \ldots \\ X_1(3)/\sigma_3 & X_2(3)/\sigma_3 & X_3(3)/\sigma_3 & \ldots \\ \ldots & \ldots & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots \end{pmatrix} \qquad (5)$$

The necessary matrix computations are readily carried out in a spreadsheet environment as well as by standard computer procedures, and they are simple extensions of the computations already required for a best fit determination.

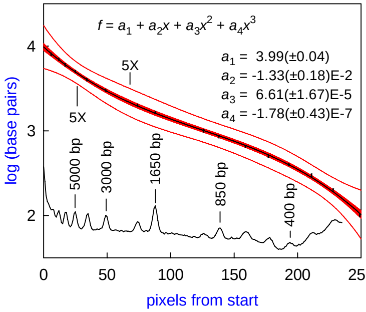
90% confidence envelopes of a quadratic fit to iron atomic absorption data (Ref. 11)

$$F \equiv 3F^{0.90}_{3,11} = 7.98$$



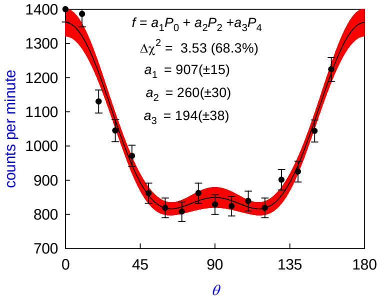
95% confidence envelopes of a cubic fit to DNA electrophoresis calibration

$$F \equiv 4F^{0.95}_{4,12} = 13.0$$



68.3% confidence envelopes for a Legendre polynomial fit to simulated nuclear counting data (Ref. 12)

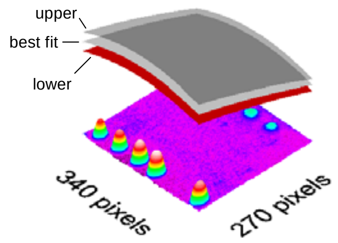
$$\Delta\chi^2 = 3.53\ (68.3\%),\ F \equiv 3F^{0.683}_{3,14} = 4.0$$



Best fit and upper and lower 68.3% confidence surfaces for background fluorescence intensities in the TLC analysis of aspirin and acetaminophen.

$$f = a_1 + a_2x + a_3y + a_4x^2 + a_5y^2 + a_6xy$$

$a_1 = 76.5(\pm 0.4)$ $a_4 = -1.79(\pm 0.16)\mathrm{E}{-4}$
$a_2 = 0.281(\pm 0.005)$ $a_5 = 6.21(\pm 0.05)\mathrm{E}{-4}$
$a_3 = 0.220(\pm 0.002)$ $a_6 = -7.9(\pm 8.4)\mathrm{E}{-6}$



Using equation (2) for the background surface fits, and replacing X_i with Q_i in (4) And (5), the equation (5) parameters are given by:

$$(\mathbf{L}) = \begin{pmatrix} 1 \\ x(i) \\ y(i) \\ x^2(i) \\ y^2(i) \\ x(i)y(i) \end{pmatrix}$$

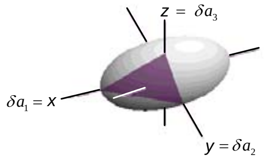
and

$$(\mathbf{Q}) = \begin{pmatrix} 1/\sigma_1 & x(1)/\sigma_1 & y(1)/\sigma_1 & x^2(1)/\sigma_1 & y^2(1)/\sigma_1 & x(1)y(1)/\sigma_1 \\ 1/\sigma_1 & x(1)/\sigma_1 & y(2)/\sigma_1 & x^2(1)/\sigma_1 & y^2(2)/\sigma_1 & x(1)y(2)/\sigma_1 \\ \ldots. & \ldots. & \ldots. & \ldots. & \ldots. & \ldots. \\ 1/\sigma_1 & x(k)/\sigma_1 & y(1)/\sigma_1 & x^2(k)/\sigma_1 & y^2(1)/\sigma_1 & x(k)y(1)/\sigma_1 \\ 1/\sigma_1 & x(k)/\sigma_1 & y(2)/\sigma_1 & x^2(k)/\sigma_1 & y^2(2)/\sigma_1 & x(k)y(2)/\sigma_1 \\ \ldots. & \ldots. & \ldots. & \ldots. & \ldots. & \ldots. \end{pmatrix}$$

Theory: A geometric depiction and vector analysis of the relationship of the $g_{max}(i)$ to the $\Delta\chi^2$ ellipsoid. For notational convenience, we use x, y and z to describe variations in the best fit parameters as indicated in the figure.

ellipse:
$0.25x^2 + 0.5y^2 + z^2 + 0.2xy + 0.5xz + 0.4yz = 3.53$

plane:
$1.70x + 2.20y + 3.00z = 7.95$



This 3D ellipsoid is defined by $\Delta\chi^2 = 3.53$ and simulates a 68.3% joint confidence surface for 3 parameters. The triangular g_{max} plane is tangent to the ellipsoid at $(x_g, y_g, z_g) = (1.89, 1.34, 0.59)$. The white vector eminates from the tangent point and is normal to both the plane and the ellipsoid. The gradient of the g plane family is given by:

$$\nabla g = \mathbf{i}\frac{\partial g}{\partial x} + \mathbf{j}\frac{\partial g}{\partial y} + \mathbf{k}\frac{\partial g}{\partial z} = \mathbf{i}X_1(i) + \mathbf{j}X_2(i) + \mathbf{k}X_3(i)$$

And the ellipsoid gradient at the tangent point is parallel to the g plane gradient, so that:

$$\mathbf{i}: \quad \frac{\partial(\Delta\chi^2)}{\partial x} = 2\alpha_{11}x_g + 2\alpha_{12}y_g + 2\alpha_{13}z_g = cX_1$$

$$\mathbf{j}: \quad \frac{\partial(\Delta\chi^2)}{\partial y} = 2\alpha_{21}x_g + 2\alpha_{22}y_g + 2\alpha_{23}z_g = cX_2$$

$$\mathbf{k}: \quad \frac{\partial(\Delta\chi^2)}{\partial z} = 2\alpha_{23}x_g + 2\alpha_{23}y_g + 2\alpha_{33}z_g = cX_3$$

A matrix analysis of the latter equation leads to equation (4)

REFERENCES

1. Massart, D.L.; Vandeginste, B.G.M.; Buydens, LM.C.; De Jong, S.; Lewi, P.J.; Smeyers-Verbeke, J. Handbook of Chemometrics and Qualimetrics: Part A, Chs. 8, 10; Elsevier, 1997.
2. Hessler, J. P.; Current, D. H.; Ogren, P. J. Comp. Phys. 1996, 10, 186.
3. Harris, D. C. Quantitative Chemical Analysis, 5th ed.; W. H. Freeman, 1999.
4. Scheffe, H. Biometrika 1953, 40, 87.
5. Dwass, M. Ann. Math. Statist. 1955, 26, 146.
6. Roy, S. N.; Bose, R. C. Ann. Math. Stat. 1953, 24, 513.
7. Miller, R. G. J. Simultaneous Statistical Inference, 2nd ed.; Springer-Verlag: New York, 1981.
8. Mandel, J.; Linnig, F. J. Anal. Chem. 1957, 29, 743.
9. Mitchell, D. G.; Mills, W. N.; Garden, J. S. Anal. Chem. 1977, 49, 1655.
10. Garden, J. S.; Mitchell, D. G.; Mills, W. N. Anal. Chem. 1980, 52, 2310.
11. Mitchell, D. G.; Garden, J. S. Talanta 1982, 29, 921.
12. Bevington, P. R.; Robinson, D. K. Data Reduction and Error Analysis for the Physical Sciences, 2nd ed.; McGraw-Hill: New York, 1993.
13. Switzer, R. L.; Garrity, L. F. Experimental Biochemistry; W. H. Freeman: New York, 1999.
14. Tellinghuisen, J. J. Mol. Spectrosc. 1996, 179, 299.
15. Chong, D. P. J. Chem. Educ. 1994, 71, 489.
16. Press, W. H.; Teukolsky, S. A.; Vetterling, W. T.; Flannery, B. P. Numerical Recipes in Fortran 77: The Art of Scientific Computing, 2nd ed.; Cambridge: New York, 1992.

ACKNOWLEDGEMENTS

This work was supported by a Howard Hughes Medical Institute grant, A National Instruments Academic Partner grant, and a Ford/Knight Endowment grant