Estimation of accuracy in large-scale linear least squares solutions

Arnold Neumaier

Fakultät für Mathematik
Universität Wien, Österreich

Lecture given on September 13, 2019 at the
MAT TRIAD 2019, Liblice, Czech Republic

These slides can be found at
`http://www.mat.univie.ac.at/~neum/ms/lsqSlides.pdf`

- Introduction
- Linear stochastic models
- Linear least squares for small systems
- Mixed linear models
- Direct methods for sparse systems
- Iterative methods for large systems

Linear stochastic models

- Expectation values
- Probability via expectation
- Covariance matrices
- Independence
- Linear models
- Linear estimators
- The Gauss–Markov theorem

Expectation values

A **random variable** is a function $x(\omega)$ depending on an experiment ω from a set Ω of conceivable experiments.

Rules for the expectation value $\overline{x} = \langle x \rangle = \mathbf{E}(x)$ of a random variable x:

(E1) $\langle 1 \rangle = 1$
(E2) $\langle \alpha x + \beta y \rangle = \alpha \langle x \rangle + \beta \langle y \rangle$
(E3) $x \geq 0 \quad \Rightarrow \quad \langle x \rangle \geq 0$
(E4) $x \geq 0,\ \langle x \rangle = 0 \quad \Rightarrow \quad x = 0$
(E5) $x_k \downarrow 0 \quad \Rightarrow \quad \langle x_k \rangle \downarrow 0$

Here α, β are real numbers,
operations and inequalities are interpreted pointwise,
and $x_k \downarrow 0$ denotes pointwise convergence from above.

Probability via expectation

The properties (E1)–(E5) can be taken as axioms for **expectation values** of an appropriate vector space of random variables.

Statements are then defined as $\{0, 1\}$-valued random variables, and **probabilities** as the expectation values of statements.

This gives a simple entrance to statistics and probability, without the need to first develop measure theory, needed only at an advanced stage.

This approach was pioneered in the excellent book

Peter Whittle, *Probability via expectation*, 3rd ed., Springer, New York 1992.

Whittle proved that this approach is equivalent to the measure theoretical definition of expectation in the traditional axiomatic approach by Kolmogorov.

Covariance matrices

The expectation (or **mean**)

$$\overline{x} = \langle x \rangle$$

of a **random vector** x is taken componentwise. The vector

$$\varepsilon := x - \overline{x}$$

is referred to as **noise vector**. The matrix

$$\operatorname{Cov}(x) := \left\langle (x - \overline{x})(x - \overline{x})^T \right\rangle = \langle \varepsilon \varepsilon^T \rangle$$

is called the **covariance matrix** of x. Every covariance matrix is symmetric and positive semidefinite, and

$$\operatorname{Cov}(Ax) = A \operatorname{Cov}(x) A^T.$$

The **variance** of x_i is

$$\operatorname{var}(x_i) = C_{ii} = \langle (x_i - \overline{x}_i)^2 \rangle = \langle \varepsilon_i^2 \rangle.$$

Its square root $\sigma_i := \sqrt{C_{ii}}$ is the **standard deviation** of x_i.

Independence

Two random vectors x, y are **independent** if

$$\langle f(x)g(y)\rangle = \langle f(x)\rangle\langle g(y)\rangle$$

for all expressions $f(x)$ in x and $g(y)$ in y.

As a consequence, x and y are **uncorrelated**, i.e., the covariance matrix of the compound vector $z = \binom{x}{y}$ is block diagonal,

$$\mathrm{Cov}(z) = \begin{pmatrix} \mathrm{Cov}(x) & 0 \\ 0 & \mathrm{Cov}(y) \end{pmatrix}.$$

Being uncorrelated is, however, significantly weaker than independence.

Linear stochastic models

A **linear stochastic model** is a relation of the form

$$y = Ax + \text{noise}(C)$$

between random vectors x and y with constant matrices A and C.

This is interpreted as the statement that the residual $\varepsilon := y - Ax$ is noise with covariance matrix C,

$$\langle \varepsilon \rangle = 0, \quad \langle \varepsilon \varepsilon^T \rangle = C.$$

As a conseqence, the means of x and y are related by $\overline{y} = A\overline{x}$, and

$$\langle \varepsilon^T M \varepsilon \rangle = \text{tr } CM$$

Linear estimators

In practice, y is a vector computable from available data while x is an unknown vector of interest to be estimated from y and an assumed linear stochastic model $y = Ax + \text{noise}(C)$.

Because of the assumed linearity, the estimators of interest are linear, i.e., of the form

$$\widetilde{x} := Sy.$$

Thus linear estimators are again random vectors.

A linear estimator $\widetilde{x}$ is **unbiased** if $\overline{\widetilde{x}} = \overline{x}$. This is a condition imposed in the absence of additional qualitative information about x. The latter would constitute the bias and can be used to regularize an otherwise ill-posed model.

The Gauss–Markov theorem

Theorem. Suppose that $y = Ax + \text{noise}(C)$. If $\begin{pmatrix} C & A \\ A^T & 0 \end{pmatrix}$ is nonsingular, there is a unique **best linear unbiased estimator (BLUE)** $\widehat{x}$ in the sense of uniform optimality:

For every linear unbiased estimator $\widetilde{x}$ and for all $a \in \mathbb{R}^n$,

$$\text{var}(a^T \widetilde{x}) \geq \text{var}(a^T \widehat{x}).$$

If C and $A^T C^{-1} A$ are nonsingular, the BLUE is given by the solution of the **normal equations**

$$A^T C^{-1} A x = A^T C^{-1} y.$$

Note that the Gauss–Markov theorem makes no assumption about the distribution of the noise beyond specifying its mean and covariance matrix.

In particular, the noise need not be Gaussian, as often assuemd.

Often in practice, C is known only inaccurately. In this case, we still get a linear unbiased estimator by solving the normal equations with a known approximation C_0 in place of the unknown C.

In place of optimality we than have an approximately optimal estimator only. When C_0 is not far from C, the variances are nearly as good as those of the the BLUE.

Linear least squares for small systems

- The linear least squares problem
- Estimating the noise level
- Estimation error for the parameters
- Prediction error for new observations
- Numerical implementation
- Regularization

The linear least squares problem

By transforming y and ε with an inverse Cholesky factor of C one can enforce that C is the identity matrix.

Then, and more generally when $C = \sigma^2 I$, corresponding to uncorrelated noise components with the same (often unknown) variance σ^2, the BLUE is given by the solution

$$\widehat{x} := (A^T A)^{-1} A^T y$$

of the normal equations

$$A^T A x = A^T y$$

for the **linear least squares problem**

$$\|Ax - y\|_2^2 = \min!$$

We concentrate on this case.

Estimating the noise level

The residual

$$\widehat{\varepsilon} := y - A\widehat{x} = y - A(A^TA)^{-1}A^Ty = \varepsilon - A(A^TA)^{-1}A^T\varepsilon$$

depends on the noise vector ε, hence is also a random vector. Its squared norm is

$$\|\widehat{\varepsilon}\|_2^2 = \widehat{\varepsilon}^*\widehat{\varepsilon} = \varepsilon^*\varepsilon - \varepsilon^*A(A^TA)^{-1}A^T\varepsilon.$$

Therefore

$$\langle\|\widehat{\varepsilon}\|_2^2\rangle = \langle\varepsilon^*\varepsilon\rangle - \sigma^2 \operatorname{tr} A(A^TA)^{-1}A^T = \sigma^2 n - \sigma^2 p,$$

Thus $\sigma^2 = \langle\|\widehat{\varepsilon}\|_2^2\rangle/(n-p)$, so that

$$\widehat{\sigma}^2 := \|\widehat{\varepsilon}\|_2^2/(n-p)$$

satisfies $\langle\widehat{\sigma}^2\rangle = \sigma^2$, hence is an unbiased estimator for the variance σ^2 from the data.

Estimation error for the parameters

The estimation error

$$\widehat{x} - \overline{x} = (A^T A)^{-1} A^T y - (A^T A)^{-1} A^T A x = (A^T A)^{-1} A^T \varepsilon$$

depends on the noise vector ε, hence is also a random vector. Its covariance matrix is

$$\mathrm{Cov}(\widehat{x} - \overline{x}) = (A^T A)^{-1} A^T \,\mathrm{Cov}(\varepsilon) A (A^T A)^{-1} = \sigma^2 (A^T A)^{-1}$$

since $\mathrm{Cov}(\varepsilon) = \sigma^2 I$. This would correspond to the root mean square error for estimating $\overline{x}$ when repeatedly solving least squares problems with the same matrix A but independent data y..

In particular, a measure for the error in the estimation of a component of $\overline{x}_i$ of $\overline{x}$ by the corresponding component $\widehat{x}_i$ of $\widehat{x}$ is its estimated standard deviation

$$\Delta\widehat{x}_i := \widehat{\sigma}\sqrt{(A^T A)^{-1}{}_{ii}}.$$

Here we replaced the often unknown model σ by its estimate $\widehat{\sigma}$.

Prediction error for new observations

Frequently one is not interested in x itself but only in the later use of the model to generate predictors for new data y_v following the model

$$y_v = A_v x + \text{noise}({\sigma_v}^2 I).$$

with a typically different coefficient matrix A_v. If we use for prediction $\widehat{x}$ in place of x we get the predictor

$$\widehat{y}_v = A_v \widehat{x}.$$

Assuming that x is not random (so that $\overline{x} = x$), the prediction error $\widehat{y}_v - y_v = A_v(\widehat{x} - x) + \text{noise}({\sigma_v}^2 I)$ has the covariance matrix

$$\text{Cov}(\widehat{y}_v - y_v) = A_v \,\text{Cov}(\widehat{x} - \overline{x}) A_v^T + {\sigma_v}^2 I = \sigma^2 A_v (A^T A)^{-1} A_v^T + {\sigma_v}^2 I.$$

Upon replacing the often unknown σ^2 and ${\sigma_v}^2$ by the estimate $\widehat{{\sigma_v}^2}$, this leads to the componentwise error estimate

$$\Delta \widehat{y}_i := \widehat{\sigma} \sqrt{\left(A_v (A^T A)^{-1} A_v^T \right)_{ii} + 1}.$$

Numerical implementation

In implementations of least squares techniques for smaller problems one generally uses a QR factorization of A into a product of an orthogonal matrix $Q \in \mathbb{R}^{n\times p}$ and an upper triangular matrix $R \in \mathbb{R}^{p\times p}$.

Then $A^TA = R^TR$, and in the well-posed case, R is nonsingular. Thus

$$(A^TA)^{-1})_{ii} = (R^{-1}R^{-T})_{ii} = \|(R^{-1})_{i:}\|_2^2,$$

$$\left(A_v(A^TA)^{-1}A_v^T\right)_{ii} = \left(A_vR^{-1}R^{-T}A_v^T\right)_{ii} = \|(A_vR^{-1})_{i:}\|_2^2,$$

giving

$$\Delta\widehat{x}_i := \widehat{\sigma}\|(R^{-1})_{i:}\|_2$$

$$\Delta\widehat{y}_i := \widehat{\sigma}\sqrt{\|A_v(R^{-1})_{i:}\|_2^2 + 1}.$$

Regularization

If R is singular, the BLUE does not exist. If R is ill-conditioned, it gives very poor estimates since R^{-1} then has (for reasonably sized A) some very large components.

Since the BLUE is only the best linear **unbiased** estimate, one can improve estimation in both cases by looking for a biased estimator.

This is done by introducing additional qualitative knowledge about x into the model.

Of course, the biased estimator is useful only when this additional qualitative knowledge is actually valid.

My survey on regularization discusses a number of different ways of posing qualitative assumptions that introducing bias in a realistic way. It is also discussed there how to handle the resulting modified least squares problems.

The only bias we consider here is that obtained by assuming that the entries of x are not large. In this case, one is lead to **ridge regression**, obtained by minimizing in place of the sum of residual squares $\|Ax - y\|_2^2$ the augmented sum $\|Ax - y\|_2^2 + \lambda\|x\|_2^2$, where λ is a regularization parameter.

In place of the standard normal equations one finds the regularized normal equations

$$(A^T A + \lambda I)x = A^T y,$$

with solution

$$\widehat{x} = (A^T A + \lambda I)^{-1} A^T y$$

depending on the regularization parameter λ.

The larger λ, the better is the condition of the linear system but the larger is also the bias. For $\lambda \to \infty$, $\widehat{x} \to 0$ and the bias wipes out all useful information. There is usually some optimal order of magnitude for λ where the tradeoff is nearly optimal.

In implementations of regularized least squares techniques for smaller problems one generally uses a singular value decomposition

$$A = U\Sigma V$$

of A into a product of two orthogonal matrices U, V and a diagonal matrix Σ containing the positive singular values σ_k in decreasing order.

Then $A^TA + \lambda I = V(\Sigma^2 + \lambda I)V^T$ and

$$\widehat{x} = (A^TA + \lambda I)^{-1}A^Ty = V\,\mathrm{Diag}\left(\frac{\sigma_k}{\sigma_k^2 + \lambda}\right)U^Ty,$$

Assuming x to be nonrandom, the error is now

$$\widehat{x} - \overline{x} = (A^TA + \lambda I)^{-1}A^Ty - x = (A^TA + \lambda I)^{-1}(A^T\varepsilon - \lambda x)$$

revealing a systematic bias that depends on the true x and cannot be estimated.

The covariance matrix $\mathrm{Cov}(\widehat{x}-\overline{x})$ of the error is now

$$(A^TA+\lambda I)^{-1}A^T\,\mathrm{Cov}(\varepsilon)A(A^TA+\lambda I)^{-1} = \sigma^2(A^TA+\lambda I)^{-1}A^TA(A^TA+\lambda I)^{-1}.$$

Thus

$$\mathrm{Cov}(\widehat{x}-\overline{x}) = \sigma^2 V\,\mathrm{Diag}\left(\frac{\sigma_k}{\sigma_k^2+\lambda}\right)^2 V^T,$$

giving (if we ignore the unknown bias and replace σ and σ_v by $\widehat{\sigma}_v$),

$$\Delta\widehat{x}_i := \widehat{\sigma}\,\sqrt{\sum_k \left(\frac{V_{ik}\sigma_k}{\sigma_k^2+\lambda}\right)^2},$$

and similarly, with $W := A_vV$,

$$\Delta\widehat{y}_i := \widehat{\sigma}\,\sqrt{\sum_k \left(\frac{W_{ik}\sigma_k}{\sigma_k^2+\lambda}\right)^2 + 1}.$$

But these formulas ignore the unknown bias, and are far too optimistic for large λ in that the error goes to zero as $\lambda \to \infty$.

Mixed linear models

- Fixed and random effects
- Large scale applications
- Mixed model equations
- The ME formulation of a mixed model
- BLUP (best linear unbiased predictor)

Fixed and random effects

A **mixed model** is a linear stochastic model of the form

$$y = X\beta + Zu + \text{noise}(D),$$

where

- y is a data vector of **records** of observed **traits**,
- β is a (fixed but unknown) coefficient vector of **fixed effects** with corresponding model matrix X,
- $u = \text{noise}(G)$ is a (random, unknown) coefficient vector of **random effects** with corresponding model matrix Z, and
- $\eta = y - X\beta - Zu = \text{noise}(D)$ is noise from observation errors and modeling errors, with covariance matrix D, uncorrelated with u.

X, Z are sparse coefficient matrices. The entries may be fixed or measured but are assumed to be deterministic. This is needed to make the stochastic model linear.

Large scale applications

I have practical experience with two classes of applications of mixed models to the analysis of huge datasets.

1. In **animal breeding**, mixed models are used to predict breeding values of animals, needed for assisting decisions about which animals to select for breeding programs. The breeding values are among the random effects. The number of parameters to be estimated depends on the population and may be in the millions.

2. In the **analysis of census data**, mixed models are used to predict among others the structure of the labor market, needed for assisting financial decisions by the government. The economic power of individuals and companies are among the random effects. The number of parameters to be estimated depends on the population of the country in which the census was done and may be in the hundreds of millions.

Mixed model equations

The mixed model equations may be rewritten as

$$y = A\beta + \text{noise}(D + ZGZ^T)$$

in terms of fixed effects β only, and is often motivated using this form. But the covariance matrix in this formulation is often large and dense, and hence expensive to use.

For solving the mixed model optimally and efficiently, Henderson 1949 discovered the **mixed model equations**

$$\begin{pmatrix} X^T D^{-1} X & X^T D^{-1} Z \\ Z^T D^{-1} X & Z^T D^{-1} Z + G^{-1} \end{pmatrix} \begin{pmatrix} \beta \\ u \end{pmatrix} = \begin{pmatrix} X^T D^{-1} y \\ Z^T D^{-1} y \end{pmatrix}.$$

for estimating only involves the inverse G^{-1}.

The mixed model equations made estimation practical at large scale since the matrix G^{-1} is usually sparse and easily computable from the practical model specifiction, without having to form G.

The vector $\widehat{u}$ (containing the estimated random effects) is usually called the **best linear unbiased predictor** (BLUP) of the random effects.

The ME formulation of a mixed model

Writing separate model equations (ME) for the measurement relation and the random effect condition gives the **ME formulation**

$$\begin{pmatrix} y \\ 0 \end{pmatrix} = \begin{pmatrix} X & Z \\ 0 & -I \end{pmatrix} \begin{pmatrix} \beta \\ u \end{pmatrix} + \text{noise} \begin{pmatrix} D & 0 \\ 0 & G \end{pmatrix}.$$

This is a linear stochastic model of the form to which the Gauss–Markov theorem applies.

As observed by Fellner 1986, Henderson's mixed model equations are just the normal equations for this model. Thus the BLUP vector $\widehat{u}$ is part of the best linear unbiased estimator $\begin{pmatrix} \widehat{\beta} \\ \widehat{u} \end{pmatrix}$ of the ME formulation of the mixed model.

For applications of the ME formulation to animal breeding, see my work with Eildert Groeneveld.

Direct methods for sparse systems

- Multifrontal solution of sparse linear systems
- Partial inversion of sparse matrices

The general theory presented applies nearly unaltered in the case of large, sparse least squares problems, provided that it is feasible to compute and store a factorization of the normal equations.

The reason is that in this case one can use the factorization to compute at little extra cost an important part of the inverse of the matrix A, namely that part that is placed on the sparsity pattern of the triangular factor.

In particular, this produces all entries B_{ik} of the inverse of the least squares matrix $B = A^T A$ with indices i, k corresponding to variables x_i, x_k appearing with nonzero coefficient in some model equation. Thus the correlations of these variables and the variances of all x_i can be found cheaply.

The large-scale formulation is most transparent in a multifrontal setting. This setting works with a blockwise factorization determined by multiple "fronts", which allow the efficient use of BLAS3 operations for dense matrix products and inverses.

Multifrontal solution of sparse linear systems

In a typical step of Gaussian elimination in block form, one factors (after a suitable symmetric reordering of rows and columns) a symmetric matrix B of the form

$$B = \begin{pmatrix} B_{JJ} & B_{JK} & 0 \\ B_{KJ} & B_{KK} & B_{KH} \\ 0 & B_{HK} & B_{HH} \end{pmatrix} = \begin{pmatrix} I & 0 & 0 \\ R_{JK}^T & I & 0 \\ 0 & 0 & I \end{pmatrix} \begin{pmatrix} D_{JJ}^{-1} & B_{JK} & 0 \\ 0 & B'_{KK} & B_{KH} \\ 0 & B_{HK} & B_{HH} \end{pmatrix}$$

using

$$D_{JJ} := B_{JJ}^{-1},\ R_{JK} = D_{JJ} B_{JK}, \tag{1}$$

$$B'_{KK} := B_{KK} - B_{KJ} R_{JK}. \tag{2}$$

(If the blocks have small size, computing the explicit block inverse involse virtually no overhead compared to the standard block elimination but helps in allowing subsequently a more efficient use of BLAS3 routines.)

Using suitable sets of indices computed in an initial symbolic factorization step, one can write the complete numerical factorzation of a sparse matrix B in terms of frontal matrices

$$\begin{pmatrix} F_{J_\nu J_\nu} & F_{J_\nu K_\nu} \\ F_{K_\nu J_\nu} & F_{K_\nu K_\nu} \end{pmatrix} = F_\nu := B(C_\nu) + \bigoplus_{\mu<\nu, \mu\sim\nu} U_\mu$$

indexed by the fronts $C_\nu := [J_\nu \;\; K_\nu]$.

% forward factorization

for $\nu = 1 : m$,

assemble F_ν using (**??**);

$D_\nu = F_{J_\nu J_\nu}^{-1}$; $R_\nu = D_\nu F_{J_\nu K_\nu}$; $U_\nu = F_{K_\nu K_\nu} - F_{K_\nu J_\nu} R_\nu$;

end;

% forward elimination

$y = b$;

for $\nu = 1 : m$,

$y_{K_\nu} = y_{K_\nu} - R_\nu^T y_{J_\nu}$

end;

% back substitution

for $\nu = m : -1 : 1$,

$x_{J_\nu} = D_\nu y_{J_\nu} - R_\nu x_{K_\nu}$;

end;

If space for the fill in is precomputed, F_{JJ} may overwrite A_{JJ}, and R and U may overwrite A_{JK}, without affecting the results.

Partial inversion of sparse matrices

Writing the inverse matrix in the form

$$M := B^{-1} = \begin{pmatrix} M_{JJ} & M_{JK} & M_{JH} \\ M_{KJ} & M_{KK} & M_{KH} \\ M_{HJ} & M_{HK} & M_{HH} \end{pmatrix},$$

we can compute the following matrix in two ways (with $*$ denoting uninteresting terms):

$$M \begin{pmatrix} I & 0 & 0 \\ R_{JK}^T & I & 0 \\ 0 & 0 & I \end{pmatrix} = \begin{pmatrix} D_{JJ} & * & * \\ 0 & \begin{bmatrix} B'_{KK} & B_{KH} \\ B_{HK} & B_{HH} \end{bmatrix}^{-1} & \end{pmatrix}$$

$$= \begin{pmatrix} M_{JJ} + M_{JK} R_{JK}^T & * & * \\ M_{KJ} + M_{KK} R_{JK}^T & M_{KK} & M_{KH} \\ * & M_{HK} & M_{HH} \end{pmatrix}.$$

Hence

$$\begin{pmatrix} B'_{KK} & B_{KH} \\ B_{HK} & B_{HH} \end{pmatrix}^{-1} = \begin{pmatrix} M_{KK} & M_{KH} \\ M_{HK} & M_{HH} \end{pmatrix}.$$

Thus this inverse is computable from the reduced problem, and once we have M_{KK}, we can find M_{JK} and M_{JJ} from

$$M_{JK} = M_{KJ}^T = -R_{JK} M_{KK}, \tag{3}$$

$$M_{JJ} = D_{JJ} - M_{JK} R_{JK}^T. \tag{4}$$

We can therefore proceed as in back substitution to recover the part of the inverse matrix covered by the fronts by what we may call **partial inversion**.

For details see my 1998 paper with Eildert Groeneveld.

% partial inversion

for $\nu = m : -1 : 1$,

$M_{J_\nu K_\nu} = -R_\nu M_{K_\nu K_\nu}$;

% fill uncomputed transposed part

$M_{K_\nu J_\nu} = M^T_{J_\nu K_\nu}$;

$M_{J_\nu J_\nu} = D_\nu - M_{J_\nu K_\nu} R^T_\nu$;

end;

Again, M_{JK} may overwrite R, and M_{JJ} may overwrite D, without changing the results. Given the factorization, the work is about twice as much as that needed for the factorization itself.

Note, however, that the inverse of a sparse matrix is generally full, hence uncomputed entries (such as M_{HJ}) are usually nonzero.

Although many nonzeros remain uncomputed (which is the reason for the efficiency of the process), the partial inverse is very important.

In many applications, and in particular for the estimation of the accuracy of the estimated parameters, the diagonal blocks of the inverse contain already all information needed, and these diagonal blocks are available from partial inversion.

The fact that part of the inverse of a sparse matrix can be computed cheaply goes back to Takahashi et al. 1973; a modern presentation can be found in Chapter 6 of Bjorck's book on numerical methods for least squares problems.

Iterative methods for large systems

(not necessarily sparse)

Iterative methods are needed for huge problems, which are so large that the calculation of a factorization is no longer feasible.

For lack of explicit inverse information, the methods treated so far are no longer applicable.

- The conjugate gradient method (CG)
- Error estimates for selected components
- k-fold cross validation

The conjugate gradient method (CG)

Linear stochastic models can be solved even for very large systems of normal equations $Bx = b$, using a routine that computes matrix vector products $q = Bp$.

One can then apply a preconditioned **conjugate gradient** method. In many cases, a few hundred iterations produce adequate accuracy, even for the largest problems.

The effort per iteration consists of

- one matrix-vector multiplication $q = Bp$,
- one application of the preconditioner (solve $Ds = r$ for s), and
- $12n + O(1)$ other operations, where n is the number of variables.

Purpose: Solves a positive definite linear system $Bx = r$

Input: r (right hand side), ε (requested relative accuracy), evaluator for $q = Bp$, solver for $Ds = r$ (preconditioner)

Requirements: B, D symmetric and positive semidefinite, D nonsingular

$p = 0;\ \ x = 0;\ \ \omega = \infty;\ \ \Delta = 0;$

while 1,

  solve $Ds = r$ for s;

  $\omega_{\text{old}} = \omega;\ \ \omega = r^T s;$

  if $\omega \leq 0$, return; end;

  $\beta = \omega/\omega_{\text{old}};\ \ p = \beta p + s;\ \ q = Bp;\ \ \alpha = p^T r/p^T q;$

  $x = x + \alpha p;\ \ \delta = \alpha^2 \omega;$

  if $\delta \leq \Delta$, return; end;

  $r = r - \alpha q;\ \ \Delta = \max(\varepsilon^2 \delta, \Delta);$

end;

Error estimates for selected components

If one needs error estimates only for a few selected components of the parameter vector, one can split the latter into a vector z of r parameters of interest and the vector x of the remaining parameters. The least squares problem then takes the form

$$\|Ax + Bz - y\|^2 = \min!$$

and the normal equations become

$$A^T Ax + A^T Bz = A^T y, \tag{5}$$

$$B^T Ax + B^T Bz = B^T y, \tag{6}$$

Insrting the solution

$$x = (A^T A)^{-1} A^T (y - Bz) \tag{7}$$

of (5) into (5) shows that z satisfies the reduced normal equations

$$(B^T B - B^T (A^T A)^{-1} A^T B)z = B^T y - B^T (A^T A)^{-1} A^T y.$$

These reduced normal equations are the normal equations for the reduced least squares problem

$$\|B'z - y'\|^2 = \min!$$

with

$$B' = B - (A^T A)^{-1} A^T B, \quad y' = y - (A^T A)^{-1} A^T y$$

obtained by substituting (7) into the original least squares problem.

The work for obtaining the reduced problem is that for linear least squares with a fixed coefficient matrix and $r + 1$ right hand sides. This can be done using any iterative method.

The reduced least squares problem has only r parameters and can be treated by a dense method, thus providing reliable error estimates.

k-fold cross validation

A general heuristics for estimating the accuracy of prediction methods is called cross validation (CV).

In k-fold cross validation, the row indices of A are randomly partitioned into $k > 1$ index sets $K_1, \ldots, K_k$ of approximately the same size. Then k different least squares problems, generated by dropping the rows indexed by one of these index sets, are solved.

The result is a sample of k CV estimates for x, from which summary statistics (mean, variance, etc.) of arbitrary functions of x can be estimated in the usual way. This provides the desired error estimates.

Compared to the BLUE, the mean of the CV estimates is slightly less accurate, by a factor of roughly $\sqrt{k/(k-1)}$ assuming a weak law of large numbers; in practice it is typically but not always significantly less, $< 5\%$ for $k \geq 4$.

One can solve in addititon the original problem and use the CV estimate only to assess the error; then no accuracy is lost.
Choosing $k = 4$ usually provides adequate error bounds, on the average within a factor of about 2 of those for the bounds provided by the standard theory.

The work for k-fold CV is essentially k times the work for solving the original least squares problem (by an arbitrary method), which is acceptable for small k when error estimation is important.

In particular, it can be used even for huge problems.

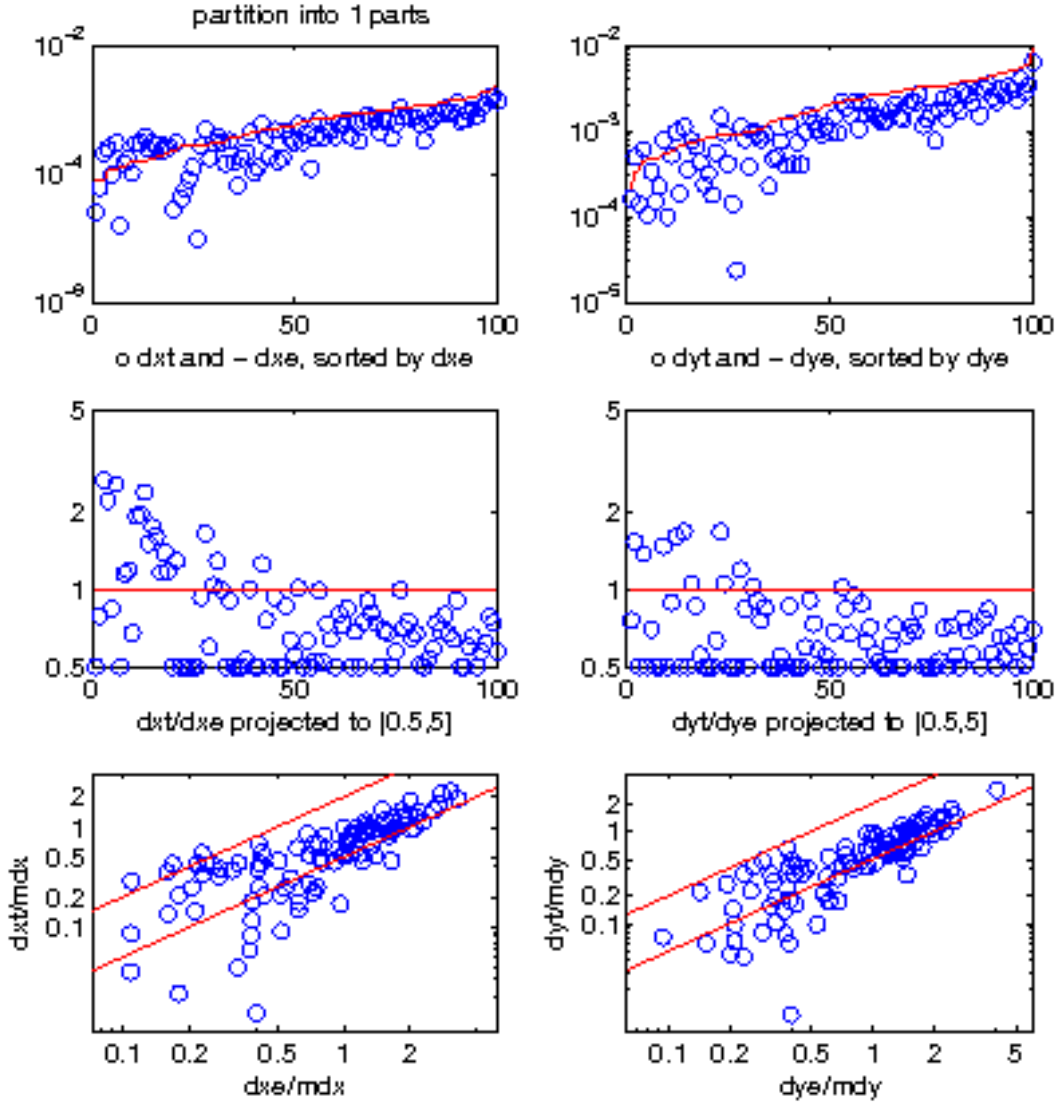
Numerical illustrations

We give illustrative simulation results for three least squares problems with $n = 5000$ data points, $p = 100$ parameters and predictions for 100 more data points.

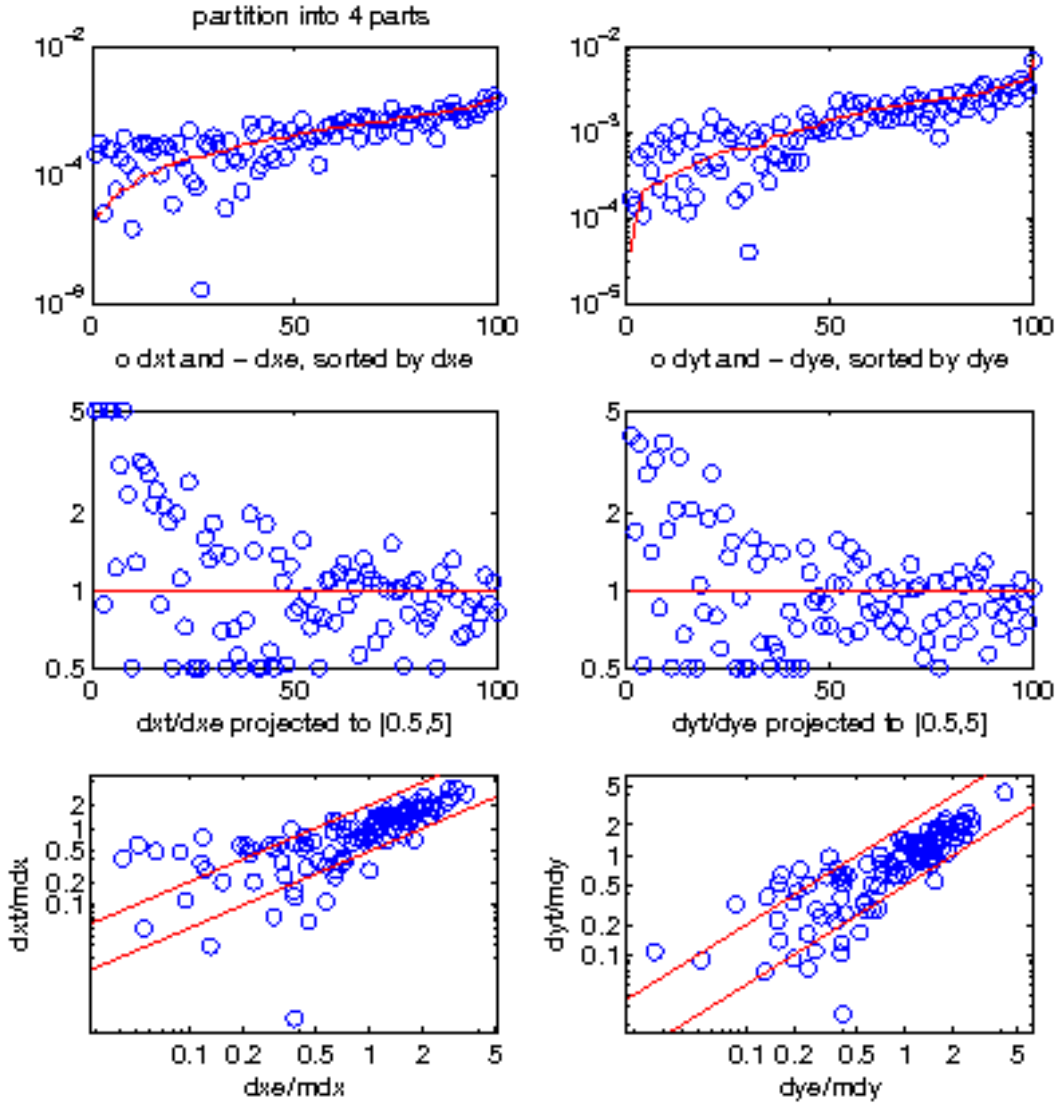
Each test problem was solved both using the standard method and using 4-fold cross validation, which is the recommended choice in the large scale case.

These figures are quite typical irrespective of the values of n and p, as long as the condition number is not too large.

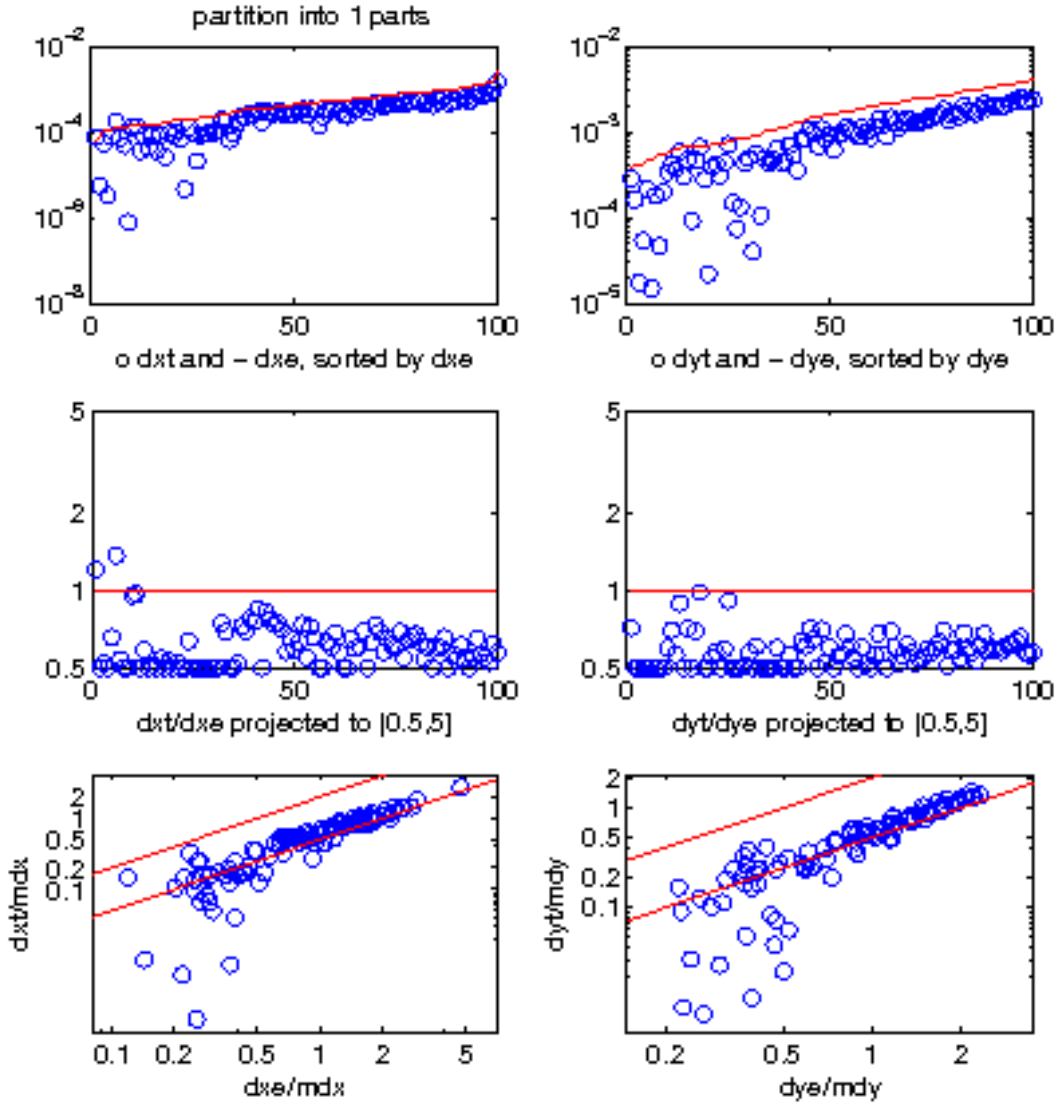
Regular case, standard estimates



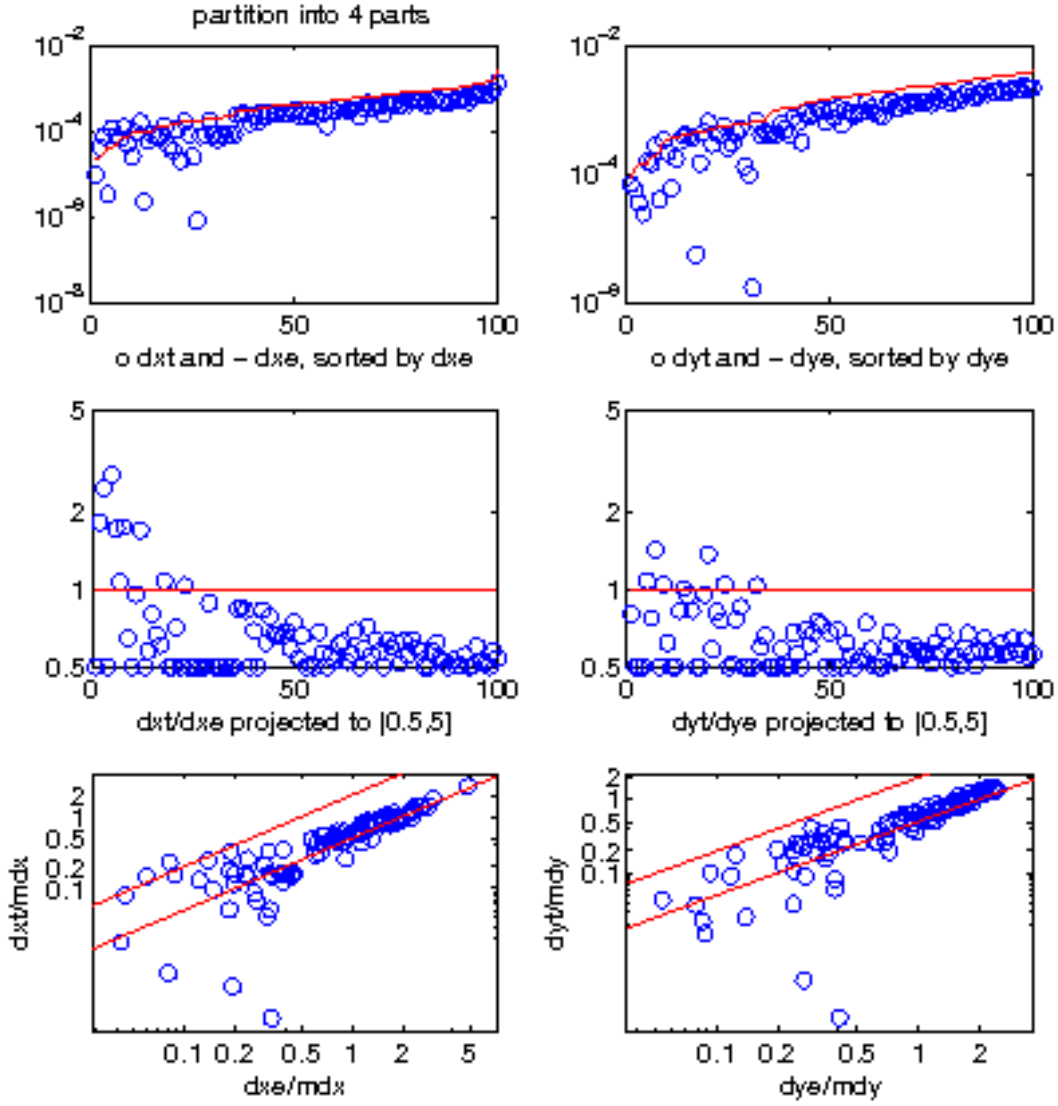
Regular case, 4-fold cross validation



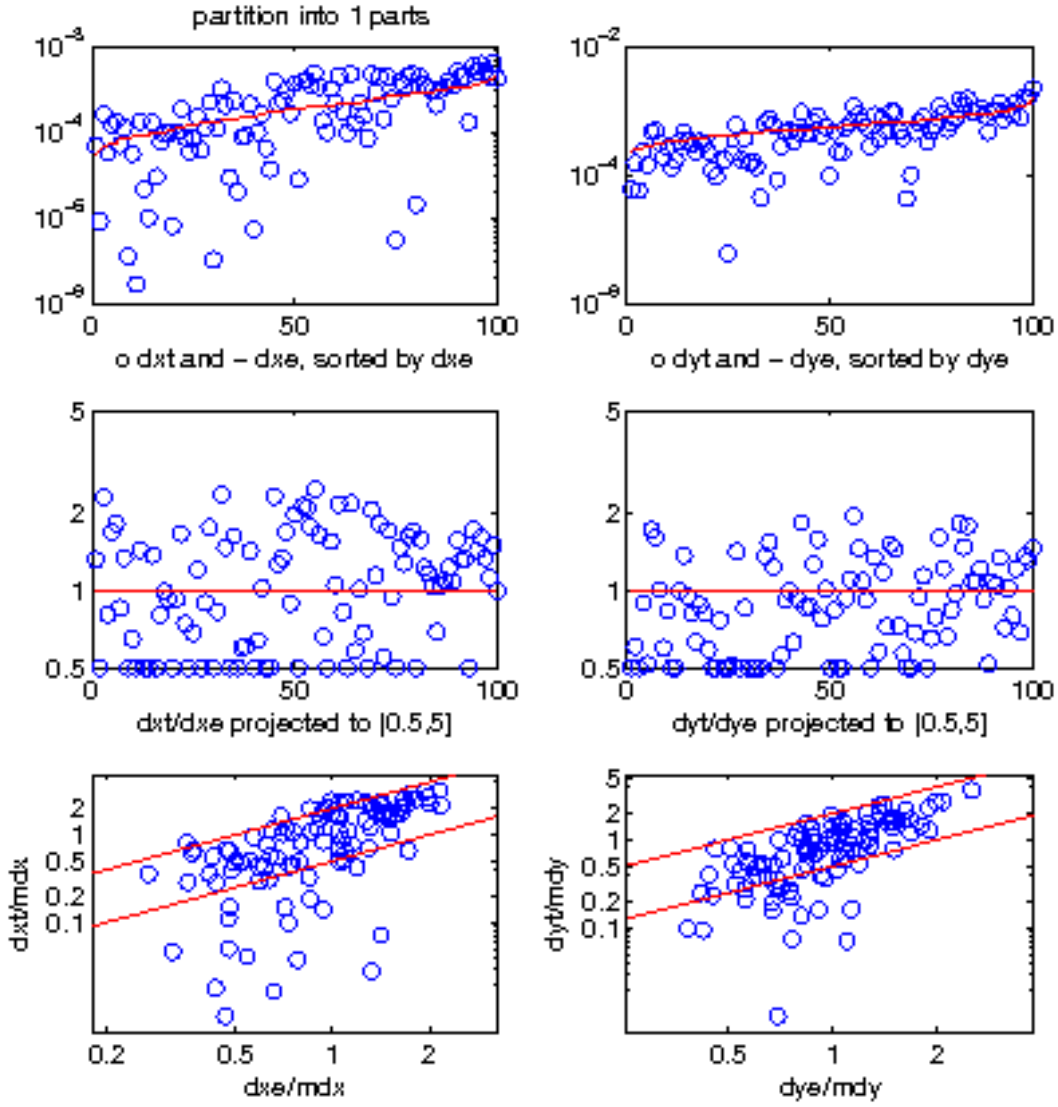
Optimistic case, standard estimates



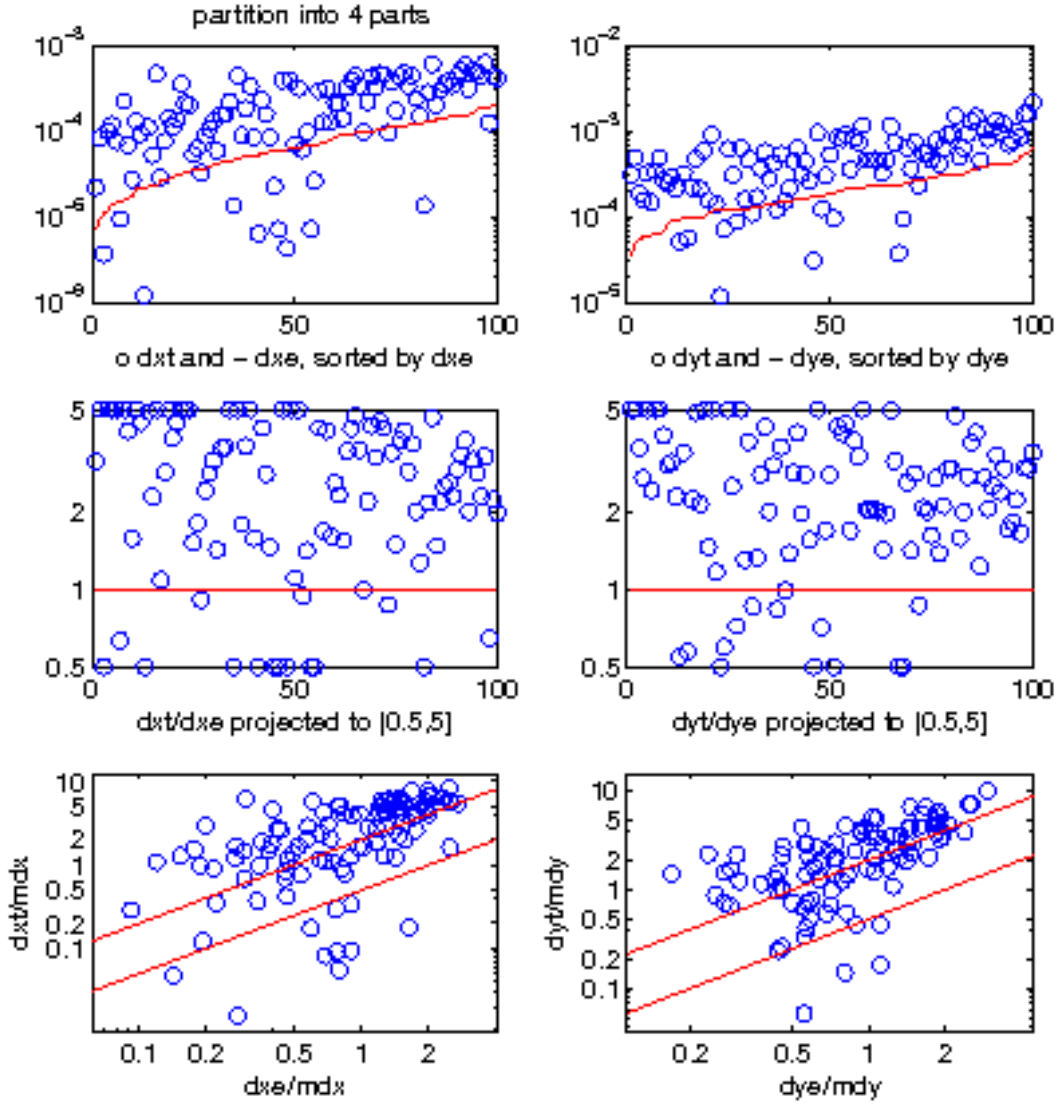
Optimistic case, 4-fold cross validation



Pessimistic case, standard estimates



Pessimistic case, 4-fold cross validation



The small n large p problem

None of the methods discussed so far works well for estimating the accuracy of the results for problems where the number n of data points is much smaller than the number p of parameters to be estimated. This situation is usually (see, e.g., Spiegelhalter 2014) referred to as the **small n large p problem** – although typically n is already large (but p is huge).

This is the case, e.g., in large mixed models for animal breeding in which both hereditary and genomic data are present.

In all small n large p problems, regularization is essential but introduces an uncontrollable amount of bias, making error estimates problematic.

There are various heuristics that can be applied to get an idea of the size of the error. The quality depends a lot on the details how these heuristics are applied.

At present I cannot recommend any of these as being reasonably reliable under reasonably general circumstances. An appropriate use of k-fold cross validation looks most promising.

Thank you for your attention!

Some references can be found on the following pages.

For a copy of the slides see

`http://www.mat.univie.ac.at/~neum/ms/lsqSlides.pdf`

References

A. Björck, Numerical methods for least squares problems, SIAM, Philadelphia 1996.

W.H. Fellner, Robust estimation of variance components, Technometrics 28 (1986), 51–60.

G. Golub and Ch.F. van Loan, Matrix computations, John Hopkins, Baltimore 1989.

E. Groeneveld and A. Neumaier,
BLUP without (inverse) relationship matrix,
Proc. World Congress Genetics Livestock Production,
vol. Theory to Application 3 (2018), 21.
`http://www.wcgalp.org/proceedings/2018/blup-without-inverse-relationship-matrix`

C.R. Henderson, Estimation of changes in herd environment, J. Dairy Sci, 32 (1949), 706–715.

A. Neumaier, Solving ill-conditioned and singular linear systems: A tutorial on regularization, SIAM Review 40 (1998), 636-666. https://www.mat.univie.ac.at/~neum/ms/regtutorial.pdf

A. Neumaier and E. Groeneveld, Restricted maximum likelihood estimation of covariances in sparse linear models, Genetics Selection Evolution 30 (1998), 3–26. http://www.mat.univie.ac.at/~neum/ms/reml.pdf

C.C. Paige and M.A. Saunders, LSQR: An algorithm for sparse linear equations and sparse least squares, ACM Transactions on Mathematical Software, 8 (1982), 43–71.

D.J. Spiegelhalter, The future lies in uncertainty, Science, 345 (2014), 264–265. http://pages.stat.wisc.edu/~wahba/spiegelhalter.science2014.pdf