Deming least-squares fit to multiple hyperplanes

Robert K. Moniot

December 29, 2006

Abstract

A method is derived to fit a set of multidimensional experimental data points having *a priori* uncertainties and possibly also covariances in all coordinates to a straight line, plane, or hyperplane of any dimensionality less than the number of coordinates. The least-squares formulation used is that of Deming, which treats all coordinates on an equal basis. Experimentalists needing to fit a linear model to data of this kind have usually performed multiple independent fits in subspaces of the full data space such that each fit has only one dependent coordinate. That procedure does not guarantee mutual consistency of the fits. The present method can be thought of as providing multiple such hyperplane fits in a single *simultaneous* and therefore consistent solution. An application to the analysis of xenon isotopes in meteorites is provided as an example.

This paper is an expanded version, with detailed derivations, of a manuscript with the same title submitted to *Applied Numerical Mathematics*.

1 Introduction

The problem posed is to fit a set of multidimensional heteroscedastic data points, i.e. data having individual *a priori* given uncertainties and possibly also covariances in all coordinates, to a hyperplane (which in two or three dimensions would be a straight line or a plane). This problem arises naturally, for instance, in analysis of noble-gas isotopes, where a measured isotopic composition is a linear combination of contributions from a number of reservoirs of fixed composition. Typically the quantity of each isotope (absolute abundance or ratio with respect to one chosen reference isotope) is associated with an uncertainty derived from internal statistics of the mass spectrometric measurement. Other experimental domains also yield data of this type, e.g. [5]. For such data, the standard method of weighted least-squares (WLS) analysis, which assumes that the independent coordinates are error-free, cannot be applied. Deming [3] provided the correct formulation of the least-squares solution of this problem. The resulting equations to be solved are non-linear. Due to the limited computing resources available at the time, Deming did not develop an exact solution to the problem. York [11, 12] was the first to provide an iterative method for solving the problem in the 2-dimensional case. Williamson [10] provided a simpler and more reliable method, also for the 2D straight-line case. Kent et al. [6] improved on these results by generalizing the method to any number of coordinates and using Newton-Raphson rather than the slower fixed-point iteration method. However, their solution is still restricted to the case of a single dependent coordinate, i.e. a hyperplane of dimensionality one less than the number of coordinates. (As will be shown, in contrast to conventional WLS, here even in the absence of covariances it is not possible to decompose a problem with multiple dependent coordinates into a set of independent problems each having just a single dependent coordinate.) For instance, their method cannot solve correctly for a straight line in three or more dimensions. They also provided only an approximate covariance matrix for the set of solution parameters.

Here the general problem of fitting data to a hyperplane with any number of dependent coordinates, i.e. of any dimensionality less than the number of coordinates, is solved, and an error propagation formula is supplied that is exact to first order in the relative variances.

An important special case is that in which the hyperplane is constrained to pass through the origin. The notation used here is able to incorporate both the constrained and unconstrained cases in a single derivation. This approach also allows the method to accommodate data sets in which some of the coordinates may be exact.

1.1   Notation

In order to define the representation of the fitted hyperplane, the coordinates of the data space are partitioned into two disjoint sets, regarded respectively as the independent coordinates (denoted by column vector $\boldsymbol{x}$) and the dependent coordinates (vector $\boldsymbol{y}$). This partitioning is arbitrary, since the Deming least-squares formulation treats all coordinates equivalently. The fitted hyperplane as a geometrical object does not depend on how the partitioning is done, although its representation does. In any case, the coordinates need to be partitioned in some way so that the representation of the hyperplane is non-redundant, that is, it has only as many parameters as degrees of freedom in the hyperplane. Then the equation of the hyperplane to be fitted can be written

$$\boldsymbol{y} = \boldsymbol{a}_0 + \mathbf{A}\boldsymbol{x} \tag{1}$$

where $\boldsymbol{a}_0$ is a vector and $\mathbf{A}$ is a matrix, which are to be determined. For the constrained case in which the hyperplane must pass through the origin, $\boldsymbol{a}_0 \equiv 0$. The constrained and unconstrained cases can be treated together by adopting the convention that for the unconstrained case, $\mathbf{A}$ is augmented with an initial column containing $\boldsymbol{a}_0$, and the vector $\boldsymbol{x}$ is correspondingly augmented with a fixed initial component equal to 1. We will place a hat on these and related quantities to signify that if dealing with the unconstrained case, they are augmented, while if it is the constrained case, they are the same as without the hat. Thus:

$$\begin{array}{cc} \textbf{Unconstrained} & \textbf{Constrained} \\ \hat{\mathbf{A}} = \left[\begin{array}{cc} \boldsymbol{a}_0 & \mathbf{A} \end{array} \right] \quad \text{and} \quad \hat{\boldsymbol{x}} = \left[\begin{array}{c} 1 \\ \boldsymbol{x} \end{array} \right] & \hat{\mathbf{A}} = \mathbf{A} \quad \text{and} \quad \hat{\boldsymbol{x}} = \boldsymbol{x}. \end{array} \tag{2}$$

The difference between the two cases then amounts to indexing columns of $\hat{\mathbf{A}}$ and elements of $\hat{\boldsymbol{x}}$ from 0 for the unconstrained case or from 1 for the constrained case. Then (1) reduces to $\boldsymbol{y} = \hat{\mathbf{A}}\hat{\boldsymbol{x}}$. Denoting the number of elements in $\hat{\boldsymbol{x}}$ by $n_{\hat{x}}$ and the number of elements in $\boldsymbol{y}$ by n_y, the size of $\hat{\mathbf{A}}$ is $n_y \times n_{\hat{x}}$. (For the constrained case $n_{\hat{x}} = n_x$, the number of independent coordinates, while for the unconstrained case $n_{\hat{x}} = n_x + 1$.)

It will be convenient to define quantities in which independent and dependent coordinates are combined. Using "blackboard bold" for this purpose,

$$\mathbb{A} = \left[\begin{array}{cc} \hat{\mathbf{A}} & -\mathbf{I}_{n_y} \end{array} \right] \quad \text{and} \quad \mathbb{x} = \left[\begin{array}{c} \hat{\boldsymbol{x}} \\ \boldsymbol{y} \end{array} \right] \tag{3}$$

so that the hyperplane equation is simply $\mathbb{A}\mathbb{x} = \mathbf{0}$. Also, it is assumed that the data points $\mathbb{x}_i$ are independent and normally distributed with known respective covariance matrices

$$\Sigma_i = \left[\begin{array}{cc} \boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_i} & \boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_i \boldsymbol{y}_i}^{\mathrm{T}} \\ \boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_i \boldsymbol{y}_i} & \boldsymbol{\Sigma}_{\boldsymbol{y}_i} \end{array} \right] \tag{4}$$

where i indexes over the set of measured data points.

1.2 Allowance for exact and inexact data

If some components of the $\mathbb{x}_i$ are exact, then Σ_i will be singular due to the presence of rows and columns that are exactly 0. (This will always be the case for the unconstrained fit, where $\boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_i}$ has an initial row and column of zeros and $\boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_i \boldsymbol{y}_i}$ has an initial column of zeros, due to the fixed initial 1 in $\hat{\boldsymbol{x}}_i$.) In the least-squares solution below, it is necessary at first to work with just the inexact parts of $\mathbb{x}_i$ and the non-singular portion of Σ_i. Partition $\hat{\boldsymbol{x}}_i$ and $\boldsymbol{y}_i$ as

$$\hat{\boldsymbol{x}}_i = \left[\begin{array}{c} \check{\boldsymbol{x}}_i \\ \tilde{\boldsymbol{x}}_i \end{array} \right] \quad \text{and} \quad \boldsymbol{y}_i = \left[\begin{array}{c} \tilde{\boldsymbol{y}}_i \\ \check{\boldsymbol{y}}_i \end{array} \right] \tag{5}$$

where the check denotes exact coordinates and the tilde denotes the coordinates with nonzero uncertainties. They are arranged so the inexact data are together in the middle of $\mathbb{x}_i$, but this is not essential; it allows a compact definition of an "extraction" matrix

$$\mathbb{E} = \left[\begin{array}{ccc} \mathbf{0} & \mathbf{I}_{\tilde{n}} & \mathbf{0} \end{array} \right] \tag{6}$$

where $\tilde{n}$ is the number of components in $\tilde{\boldsymbol{x}}_i$ and $\tilde{\boldsymbol{y}}_i$, so that the portion of $\mathbb{x}_i$ with uncertainties is

$$\mathbb{E}\mathbb{x}_i = \left[\begin{array}{c} \tilde{\boldsymbol{x}}_i \\ \tilde{\boldsymbol{y}}_i \end{array} \right] \doteq \tilde{\mathbb{x}}_i. \tag{7}$$

Other arrangements of the exact and inexact data can be handled by permuting the columns of $\mathbb{E}$ to match; all the results below involving $\mathbb{E}$ are unchanged if this is done. Then the non-singular portion of the covariance matrix, that is, the portion that applies only to $\tilde{\boldsymbol{x}}_i$ and $\tilde{\boldsymbol{y}}_i$, is

$$\tilde{\Sigma}_i = \mathbb{E}\, \Sigma_i\, \mathbb{E}^{\mathrm{T}}. \tag{8}$$

2 Least-squares solution

2.1 Deming least-squares criterion

Deming's [3] formulation of the least-squares criterion for a fit to data with uncertainties in the independent as well as dependent coordinates begins by introducing "adjusted" data points $\mathbb{\mu}_i$, which can be regarded as the means of the respective ideal populations from which the $\mathbb{x}_i$ are drawn, where again i indexes over the set of measurements. By hypothesis the adjusted values obey the equation of the hyperplane exactly:

$$\mathbb{A}\mathbb{\mu}_i = \mathbf{0}, \quad \forall i. \tag{9}$$

Then, on the assumption of normally distributed errors, the log likelihood of the set of measurements is (suppressing the normalization constant)

$$\begin{aligned} L &= -\tfrac{1}{2} \sum_i (\tilde{\mathbb{x}}_i - \tilde{\mathbb{\mu}}_i)^{\mathrm{T}} \tilde{\Sigma}_i^{-1} (\tilde{\mathbb{x}}_i - \tilde{\mathbb{\mu}}_i) \\ &= -\tfrac{1}{2} \sum_i (\mathbb{x}_i - \mathbb{\mu}_i)^{\mathrm{T}} \mathbb{E}^{\mathrm{T}} \tilde{\Sigma}_i^{-1} \mathbb{E} (\mathbb{x}_i - \mathbb{\mu}_i). \end{aligned} \tag{10}$$

This quantity is to be maximized subject to the constraint (9). Note that the components of $\mathbb{\mu}_i$ that are eliminated by $\mathbb{E}$ are defined to be equal to the corresponding (exact) components of $\mathbb{x}_i$, and so they do not contribute to L, nor do they enter into the maximization.

First L is maximized with respect to the μ_i while holding $\mathbb{A}$ fixed. Take the inner product of (9) for each i with a vector $\boldsymbol{\lambda}_i$ of Lagrange multipliers and add to (10), then differentiate the result with respect to the μ_i:

$$(\varkappa_i - \mu_i)^{\mathrm{T}} \mathbb{E}^{\mathrm{T}} \tilde{\Sigma}_i^{-1} \mathbb{E}\, \delta\mu_i + \boldsymbol{\lambda}_i^{\mathrm{T}} \mathbb{A}\, \delta\mu_i = 0, \quad \forall i. \tag{11}$$

Since the components of $\delta\mu_i$ corresponding to the exact components of $\varkappa_i$ are 0 by definition, $\mathbb{A}\,\delta\mu_i = \mathbb{A}\,\mathbb{E}^{\mathrm{T}}\mathbb{E}\,\delta\mu_i = \mathbb{A}\,\mathbb{E}^{\mathrm{T}}\,\delta\tilde{\mu}_i$ so (11) can be written in terms of only the variable components of $\varkappa_i$ and μ_i as

$$(\tilde{\varkappa}_i - \tilde{\mu}_i)^{\mathrm{T}} \tilde{\Sigma}_i^{-1} \delta\tilde{\mu}_i + \boldsymbol{\lambda}_i^{\mathrm{T}} \mathbb{A}\,\mathbb{E}^{\mathrm{T}}\, \delta\tilde{\mu}_i = 0, \quad \forall i. \tag{12}$$

Equating coefficients of $\delta\tilde{\mu}_i$ to 0 yields

$$\tilde{\mu}_i = \tilde{\varkappa}_i + \tilde{\Sigma}_i\, \mathbb{E}\, \mathbb{A}^{\mathrm{T}} \boldsymbol{\lambda}_i, \quad \forall i. \tag{13}$$

Multiplication on the left by $\mathbb{E}^{\mathrm{T}}$ expands a vector like $\tilde{\varkappa}_i$ to full size while padding with zeros at the positions of exact components. Therefore

$$\mu_i = \varkappa_i + \mathbb{E}^{\mathrm{T}} \tilde{\Sigma}_i\, \mathbb{E}\, \mathbb{A}^{\mathrm{T}} \boldsymbol{\lambda}_i = \varkappa_i + \Sigma_i\, \mathbb{A}^{\mathrm{T}} \boldsymbol{\lambda}_i, \quad \forall i \tag{14}$$

for all the components of μ_i. Inserting this result into the constraint equation (9) and solving,

$$\boldsymbol{\lambda}_i = -\left(\mathbb{A}\Sigma_i\mathbb{A}^{\mathrm{T}}\right)^{-1} \mathbb{A}\,\varkappa_i, \quad \forall i. \tag{15}$$

Since $\mathbb{A}\Sigma_i\mathbb{A}^{\mathrm{T}} = \mathbb{A}\,\mathbb{E}^{\mathrm{T}}\tilde{\Sigma}_i\,\mathbb{E}\,\mathbb{A}^{\mathrm{T}}$, and $\tilde{\Sigma}_i$ is non-singular, it follows that the inverse of $\mathbb{A}\Sigma_i\mathbb{A}^{\mathrm{T}}$ exists provided $\mathrm{rank}(\mathbb{A}\,\mathbb{E}^{\mathrm{T}}) = n_y$. This requires $\tilde{n} \geq n_y$, a condition that is normally satisfied for the problems addressed here: for instance it is sufficient to require that all of the dependent variables have uncertainties.

It is worth noting that the terms $\mathbb{A}\Sigma_i\mathbb{A}^{\mathrm{T}}$ are simply the inverse of Deming's weights (see e.g. p. 179 of [3]). Thus define

$$\mathbf{W}_i^{-1} \doteq \mathbb{A}\Sigma_i\mathbb{A}^{\mathrm{T}}. \tag{16}$$

Defining also the conventional residuals $\boldsymbol{r}_i \doteq \mathbb{A}_i\varkappa_i$ gives $\boldsymbol{\lambda}_i = -\mathbf{W}_i\boldsymbol{r}_i$. Then the Deming residuals are given by

$$\varkappa_i - \mu_i = \Sigma_i\mathbb{A}^{\mathrm{T}}\mathbf{W}_i\boldsymbol{r}_i, \quad \forall i, \tag{17}$$

which allows (10) to be rewritten compactly as

$$L_{\mathrm{p}} = -\tfrac{1}{2} \sum_i \boldsymbol{r}_i^{\mathrm{T}} \mathbf{W}_i \boldsymbol{r}_i. \tag{18}$$

The subscript p signifies that this is now the profile log likelihood, i.e. a function only of $\mathbb{A}$. This has the same form as for conventional WLS, but differs from it in that the weights $\mathbf{W}_i$ are not fixed *a priori,* but depend on $\mathbb{A}$. Note that if the $\mathbf{W}_i$ are diagonal, (18) can be decomposed into a set of sums for each dependent coordinate that can be maximized independently. But the $\mathbf{W}_i$ are not diagonal unless the independent coordinates are exact ($\boldsymbol{\Sigma}_{\boldsymbol{x}_i} = \mathbf{0}$) and the dependent coordinates are uncorrelated ($\boldsymbol{\Sigma}_{\boldsymbol{y}_i}$ diagonal), i.e. for conventional WLS with uncorrelated data.

2.2 Quasi-normal equations

Now L_{p} is maximized with respect to $\mathbb{A}$. Since part of $\mathbb{A}$ is fixed (the identity matrix), it needs to be unpacked to extract the unknown portion $\hat{\mathbf{A}}$. This gives

$$\begin{aligned}
\boldsymbol{r}_i &= \hat{\mathbf{A}}\hat{\boldsymbol{x}}_i - \boldsymbol{y}_i && (19)\\
\mathbf{W}_i^{-1} &= \hat{\mathbf{A}}\boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_i}\hat{\mathbf{A}}^{\mathrm{T}} - \hat{\mathbf{A}}\boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_i\boldsymbol{y}_i}^{\mathrm{T}} - \boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_i\boldsymbol{y}_i}\hat{\mathbf{A}}^{\mathrm{T}} + \boldsymbol{\Sigma}_{\boldsymbol{y}_i}. && (20)
\end{aligned}$$

In fact the hats can be removed from all quantities in (20) because for the unconstrained case, the initial column of $\hat{\mathbf{A}}$ (the vector $\boldsymbol{a}_0$ defined in section 1.1) does not contribute to $\mathbf{W}_i^{-1}$ due to the zeros in $\boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_i}$ and $\boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_i \boldsymbol{y}_i}$. The hats are left on for now to maintain uniform treatment in the rest of the derivation. Similarly unpack

$$\mu_i = \left[\begin{array}{c} \hat{\boldsymbol{\mu}}_i \\ \boldsymbol{\eta}_i \end{array} \right] \tag{21}$$

with $\hat{\boldsymbol{\mu}}_i$ having $n_{\hat{x}}$ components and $\boldsymbol{\eta}_i$ having n_y components. Minimizing (18) with respect to $\hat{\mathbf{A}}$,

$$\delta L_{\mathrm{p}} \quad = \quad -\sum_i \left[\boldsymbol{r}_i^{\mathrm{T}} \mathbf{W}_i \, \delta \boldsymbol{r}_i + \tfrac{1}{2} \boldsymbol{r}_i^{\mathrm{T}} \, \delta \mathbf{W}_i \, \boldsymbol{r}_i \right] = \mathbf{0}. \tag{22}$$

From (19) and (20)

$$\delta \boldsymbol{r}_i \quad = \quad \delta \hat{\mathbf{A}} \, \hat{\boldsymbol{x}}_i \tag{23}$$

$$\delta \mathbf{W}_i \quad = \quad -\mathbf{W}_i \, \delta \mathbf{W}_i^{-1} \, \mathbf{W}_i = -\mathbf{W}_i \left[\delta \hat{\mathbf{A}} \, \hat{\boldsymbol{\Xi}}_i^{\mathrm{T}} + \hat{\boldsymbol{\Xi}}_i \, \delta \hat{\mathbf{A}}^{\mathrm{T}} \right] \mathbf{W}_i \tag{24}$$

where $\hat{\boldsymbol{\Xi}}_i \doteq \hat{\mathbf{A}} \boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_i} - \boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_i \boldsymbol{y}_i}$. Inserting these expressions into (22) and using

$$\hat{\boldsymbol{\mu}}_i = \hat{\boldsymbol{x}}_i - \hat{\boldsymbol{\Xi}}_i^{\mathrm{T}} \mathbf{W}_i \boldsymbol{r}_i \tag{25}$$

from unpacking (17), one obtains

$$\delta L_{\mathrm{p}} \quad = \quad -\sum_i \boldsymbol{r}_i^{\mathrm{T}} \mathbf{W}_i \, \delta \hat{\mathbf{A}} \, \hat{\boldsymbol{\mu}}_i = \mathbf{0}. \tag{26}$$

Equating the coefficient of each element of $\delta \hat{\mathbf{A}}$ to $\mathbf{0}$ yields $\sum_i \mathbf{W}_i \, \boldsymbol{r}_i \hat{\boldsymbol{\mu}}_i^{\mathrm{T}} = \mathbf{0}$ or

$$\sum_i \mathbf{W}_i \, \hat{\mathbf{A}} \, \hat{\boldsymbol{x}}_i \hat{\boldsymbol{\mu}}_i^{\mathrm{T}} = \sum_i \mathbf{W}_i \, \boldsymbol{y}_i \hat{\boldsymbol{\mu}}_i^{\mathrm{T}}. \tag{27}$$

This system closely resembles the normal equations of the conventional WLS problem, but using the Deming weights $\mathbf{W}_i$ and replacing one instance of $\hat{\boldsymbol{x}}_i$ on each side by the adjusted value $\hat{\boldsymbol{\mu}}_i$. In fact, if only the dependent variables have uncertainties, i.e. $\boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_i} = \mathbf{0}$ and $\boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_i \boldsymbol{y}_i} = \mathbf{0}$, then $\mathbf{W}_i^{-1} = \boldsymbol{\Sigma}_{\boldsymbol{y}_i}$ and $\hat{\boldsymbol{\mu}}_i = \hat{\boldsymbol{x}}_i$, $\forall i$. Then (27) becomes the conventional WLS normal equations. However, in the general case these equations are not normal, though they will be close to normal if the Deming residuals are small, so I will call them quasi-normal.

It is worth digressing briefly to consider whether it is possible to avoid solving the quasi-normal equations, which like their conventional cousins tend to be ill-conditioned. The preferred method of solving the conventional WLS problem is to rewrite (18) as a norm of residuals:

$$-2L_{\mathrm{p}} = \sum_i (\hat{\mathbf{A}} \hat{\boldsymbol{x}}_i - \boldsymbol{y}_i)^{\mathrm{T}} \mathbf{W}_i (\hat{\mathbf{A}} \hat{\boldsymbol{x}}_i - \boldsymbol{y}_i) \quad \to \quad \| \hat{\mathbf{A}} \mathbf{X} - \mathbf{Y} \|_2^2 \tag{28}$$

where $\mathbf{X}$ and $\mathbf{Y}$ are matrices whose columns are the data vectors $\hat{\boldsymbol{x}}_i$ and $\boldsymbol{y}_i$ respectively, scaled by the square roots of the weights. This norm is minimized by $\mathbf{A} = \mathbf{Y}\mathbf{X}^+$ where $\mathbf{X}^+$ is the generalized inverse of $\mathbf{X}$. (See, for instance Golub and Van Loan [4]. Their proof is for $n_y = 1$ but it generalizes readily to $n_y \geq 1$.) However, in the present case besides the fact that the $\mathbf{W}_i$ are matrices and so cannot in general be used to scale the $\hat{\boldsymbol{x}}_i$, the approach fails because the $\mathbf{W}_i$ are not fixed constants, but depend on the solution $\hat{\mathbf{A}}$. Even if it were possible to rewrite L_{p} in terms of a norm as in (28), minimizing this norm would be equivalent to solving the conventional (not quasi) normal equations with the $\mathbf{W}_i$ evaluated at the solution. It would therefore not give the Deming least-squares solution.

2.3 Solution of the quasi-normal equations

2.3.1 Fixed-point iteration

The quasi-normal equations (27) are non-linear, and do not have a closed-form solution. Therefore an iterative solution is necessary. Although fixed-point iteration has only linear convergence, it is worth examining because as formulated below it is very stable and tends to reach a good approximate solution quickly. Thus it can provide a good starting point for the quadratically convergent Newton-Raphson iteration which is derived later.

Noting that the dependence of $\hat{\boldsymbol{\mu}}_i$ on $\hat{\mathbf{A}}$ is weak, and that (27) is not very sensitive to changes in the $\mathbf{W}_i$ with $\hat{\mathbf{A}}$ (proportional changes would cancel out altogether), a fixed-point iteration can be based on treating the $\hat{\boldsymbol{\mu}}_i$ and $\mathbf{W}_i$ as constants. The system (27) can be vectorized as

$$\left(\sum_i \hat{\boldsymbol{\mu}}_i \hat{\boldsymbol{x}}_i^{\mathrm{T}} \otimes \mathbf{W}_i \right) \operatorname{vec} \hat{\mathbf{A}} = \sum_i \hat{\boldsymbol{\mu}}_i \otimes \mathbf{W}_i \boldsymbol{y}_i \qquad (29)$$

In this notation the vectorization operator vec yields a column vector formed by concatenating successive columns of the matrix. The symbol $\otimes$ denotes the Kronecker product. See the Appendix for a summary of the relevant properties of the Kronecker product. The iteration proceeds by evaluating the two sums using the current guess for $\hat{\mathbf{A}}$, solving for the new $\hat{\mathbf{A}}$, and repeating. Each iteration requires solving an unsymmetric system of $n_{\hat{x}} n_y$ equations for the $n_{\hat{x}} n_y$ elements of $\hat{\mathbf{A}}$. Provided the $\boldsymbol{\Sigma}_{\boldsymbol{y}_i}$ are not singular, an initial guess of $\hat{\mathbf{A}} = \mathbf{0}$ works well: if the Σ_i are diagonal then with this starting point the first iteration gives the result of using WLS with weights of $\boldsymbol{\Sigma}_{\boldsymbol{y}_i}^{-1}$, which should be close to the Deming fit. Commonly the accuracy of the solution improves by more than one decimal digit per iteration, so it is usually worth doing one or two more iterations to provide the Newton-Raphson method with a very good starting point from which it can converge to machine precision in just a few iterations.

Since the Deming weights are symmetric positive definite, they can be Cholesky factored as $\mathbf{W}_i = \mathbf{G}_i \mathbf{G}_i^{\mathrm{T}}$ so the quasi-normal matrix in parentheses in (29) can be factored as $\sum_i (\hat{\boldsymbol{\mu}}_i \otimes \mathbf{G}_i)(\hat{\boldsymbol{x}}_i \otimes \mathbf{G}_i)^{\mathrm{T}}$. Since the adjusted points $\hat{\boldsymbol{\mu}}_i$ should be close to the $\hat{\boldsymbol{x}}_i$, this shows that in practice this matrix will usually be positive definite (provided there are sufficient independent data points for it to be of full rank) but for badly scattered data, large uncertainties and covariances, or inappropriate guesses for $\hat{\mathbf{A}}$ this cannot be guaranteed. Therefore a general-matrix solver should be used to solve (29).

2.3.2 Newton-Raphson iteration

Defining

$$\hat{\mathbf{F}}(\hat{\mathbf{A}}) \doteq \sum_i \mathbf{W}_i \left(\hat{\mathbf{A}} \hat{\boldsymbol{x}_i} - \boldsymbol{y}_i \right) \hat{\boldsymbol{\mu}}_i^{\mathrm{T}} \qquad (30)$$

then by (27) the solution to the least-squares problem is given by $\hat{\mathbf{F}}(\hat{\mathbf{A}}) = \mathbf{0}$. This equation can be solved using the multivariate Newton-Raphson method. (See, e.g. Baldick [1]. For $n_y = 1$ the resulting iteration is the same as the method of Kent et al. [6].) This method requires evaluating the Jacobian matrix of $\hat{\mathbf{F}}$ with respect to $\hat{\mathbf{A}}$.

Since various conventions for matrix derivatives are found in the literature, it is necessary to define the notation used here. For a vector $\boldsymbol{f}$ that is a function of a vector $\boldsymbol{x}$, $\mathrm{d}\boldsymbol{f}/\mathrm{d}\boldsymbol{x}$ is defined to be a matrix with $\mathrm{d}f_i/\mathrm{d}x_j$ in the i-th row and j-th column. In this convention, the gradient of a scalar is a row vector. Then for a matrix $\mathbf{F}$ that is a function of a matrix $\mathbf{X}$, $\mathrm{d}\mathbf{F}/\mathrm{d}\mathbf{X} \doteq \mathrm{d}\operatorname{vec}\mathbf{F}/\mathrm{d}\operatorname{vec}\mathbf{X}$. (Writing a matrix derivative as a ratio of differentials is

deprecated by some authors, e.g. [8], for good reasons, but this notation is retained here because it makes many of the following expressions more readable as analogs of their scalar counterparts.)

With this definition of the matrix derivative, we solve for the Jacobian matrix $\hat{\mathbf{J}} \doteq \partial \hat{\mathbf{F}}/\partial \hat{\mathbf{A}}$. Details are given in the Appendix, Sec. 8.1. The result is

$$
\hat{\mathbf{J}} = \sum_i \left\{ \hat{\boldsymbol{\mu}}_i \hat{\boldsymbol{\mu}}_i^{\mathrm{T}} \otimes \mathbf{W}_i - \boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_i} \otimes \mathbf{W}_i \boldsymbol{r}_i \boldsymbol{r}_i^{\mathrm{T}} \mathbf{W}_i + \hat{\boldsymbol{\Xi}}_i^{\mathrm{T}} \mathbf{W}_i \hat{\boldsymbol{\Xi}}_i \otimes \mathbf{W}_i \boldsymbol{r}_i \boldsymbol{r}_i^{\mathrm{T}} \mathbf{W}_i \right.
$$
$$
\left. - \mathbf{W}_i \hat{\boldsymbol{\Xi}}_i \bowtie \mathbf{W}_i \boldsymbol{r}_i \hat{\boldsymbol{\mu}}_i^{\mathrm{T}} - \mathbf{W}_i \boldsymbol{r}_i \hat{\boldsymbol{\mu}}_i^{\mathrm{T}} \bowtie \mathbf{W}_i \hat{\boldsymbol{\Xi}}_i \right\}. \tag{31}
$$

Because $(\mathbf{A} \otimes \mathbf{B})^{\mathrm{T}} = \mathbf{A}^{\mathrm{T}} \otimes \mathbf{B}^{\mathrm{T}}$ and $(\mathbf{A} \bowtie \mathbf{B})^{\mathrm{T}} = \mathbf{B} \bowtie \mathbf{A}$, $\hat{\mathbf{J}}$ in (31) is manifestly symmetric. It can also be shown to be positive definite, provided the residuals are small compared to the data values. The proof is as follows. Let $\hat{\boldsymbol{\Gamma}}_i$ denote the Cholesky factorization of $\boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_i}$, i.e. $\hat{\boldsymbol{\Gamma}}_i \hat{\boldsymbol{\Gamma}}_i^{\mathrm{T}} = \boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_i}$, and define

$$
\hat{\mathbf{M}}_i \doteq \hat{\boldsymbol{\mu}}_i \otimes \mathbf{I}_{n_y} \tag{32}
$$
$$
\hat{\mathbf{E}}_i \doteq \mathbf{W}_i \boldsymbol{r}_i \bowtie \hat{\boldsymbol{\Xi}}_i \tag{33}
$$
$$
\hat{\mathbf{S}}_i \doteq \hat{\boldsymbol{\Gamma}}_i \otimes \mathbf{W}_i \boldsymbol{r}_i \tag{34}
$$

(The notation reflects the identification of $\hat{\mathbf{M}}_i$ as a kind of inflated $\hat{\boldsymbol{\mu}}_i$, and $\hat{\mathbf{E}}_i$ as a relative of $\hat{\boldsymbol{\epsilon}}_i \doteq \hat{\boldsymbol{x}}_i - \hat{\boldsymbol{\mu}}_i = \hat{\boldsymbol{\Xi}}_i^{\mathrm{T}} \mathbf{W}_i \boldsymbol{r}_i$. In the $n_y = 1$ case in fact $\hat{\mathbf{M}}_i = \hat{\boldsymbol{\mu}}_i$ and $\hat{\mathbf{E}}_i = \hat{\boldsymbol{\epsilon}}_i$.) Then

$$
\hat{\mathbf{J}} = \sum_i \left\{ \hat{\mathbf{M}}_i \mathbf{W}_i \hat{\mathbf{M}}_i^{\mathrm{T}} + \hat{\mathbf{E}}_i \mathbf{W}_i \hat{\mathbf{E}}_i^{\mathrm{T}} - \hat{\mathbf{M}}_i \mathbf{W}_i \hat{\mathbf{E}}_i^{\mathrm{T}} - \hat{\mathbf{E}}_i \mathbf{W}_i \hat{\mathbf{M}}_i^{\mathrm{T}} - \hat{\mathbf{S}}_i \hat{\mathbf{S}}_i^{\mathrm{T}} \right\}
$$
$$
= \sum_i \left\{ (\hat{\mathbf{M}}_i - \hat{\mathbf{E}}_i) \mathbf{G}_i \mathbf{G}_i^{\mathrm{T}} (\hat{\mathbf{M}}_i - \hat{\mathbf{E}}_i)^{\mathrm{T}} - \hat{\mathbf{S}}_i \hat{\mathbf{S}}_i^{\mathrm{T}} \right\} \tag{35}
$$

where $\mathbf{G}_i$ is the Cholesky factorization of $\mathbf{W}_i$ as in Sec. 2.3.1. Since the $\hat{\mathbf{M}}_i$ are of the order of magnitude of the data, while $\hat{\mathbf{E}}_i$ and $\hat{\mathbf{S}}_i$ are proportional to the residuals, the terms in the sum will all be positive provided the residuals are small relative to the data.

Thus $\hat{\mathbf{J}}$ may not be positive definite if $\hat{\mathbf{A}}$ may have any value. At convergence, however, where $\hat{\mathbf{A}}$ minimizes L_{p}, $\hat{\mathbf{J}}$ must be positive definite since it is the Hessian matrix of L_{p} with respect to $\hat{\mathbf{A}}$.

3 Error analysis

3.1 Goodness of fit

Having solved for $\hat{\mathbf{A}}$, the goodness of fit can be assessed. The assumption of normally distributed errors implies (as shown by Deming [3]) that the sum in (10) gives the variance of the fit, which approximates χ^2. This variance is more conveniently calculated using the equivalent sum in (18):

$$
\chi^2 \approx -2L_{\mathrm{p}} = \sum_i \boldsymbol{r}_i^{\mathrm{T}} \mathbf{W}_i \boldsymbol{r}_i \tag{36}
$$

If the data fit the assumed form of a hyperplane, then χ^2 should be approximately equal to the number of degrees of freedom, $\nu = m n_y - n_{\hat{x}} n_y = (m - n_{\hat{x}}) n_y$, where m is the number of data points. This number is derived by considering that each point contributes only n_y degrees of freedom, one for each dependent coordinate, since the location of the independent

coordinates is arbitrary. Then one subtracts the number of parameters in $\hat{\mathbf{A}}$. The reduced χ^2, $\chi^2_\nu \doteq \chi^2/\nu$, should therefore be approximately equal to 1. If χ^2_ν is significantly larger than 1, then either the variances and covariances of the data points have been underestimated *a priori,* or else the data do not fit the model well but scatter off the hyperplane. The latter explanation could be due to the data requiring a higher-dimensionality hyperplane than assumed (e.g. in the context of isotopic ratios, if the samples are mixtures of a larger number of reservoirs than assumed), or due to a failure of the linear form (e.g. data that should be fit to a nonlinear function). If χ^2_ν is significantly smaller than 1, then either the variances and covariances of the data points have been overestimated, or the data points actually fit a lower-dimensionality hyperplane, so that the unneeded degrees of freedom allow a specious improvement in the closeness of the data to the hyperplane.

3.2 Covariance matrix of $\hat{\mathbf{A}}$

For the conventional WLS case, the Gauss-Markov theorem implies that the covariance matrix of the elements of $\hat{\mathbf{A}}$ is given by the inverse of the Hessian matrix of L_p with respect to $\hat{\mathbf{A}}$. This matrix is the same as $\hat{\mathbf{J}}$, the Jacobian matrix of $\hat{\mathbf{F}}$ as calculated above, so this suggests

$$\mathbf{\Sigma}_{\hat{\mathbf{A}}} \approx \hat{\mathbf{J}}^{-1}. \tag{37}$$

This is an $n_{\hat{x}} n_y \times n_{\hat{x}} n_y$ matrix whose i, j element is $\mathrm{cov}(a_i, a_j)$ where $\boldsymbol{a} = \mathrm{vec}\,\hat{\mathbf{A}}$. This expression for $\mathbf{\Sigma}_{\hat{\mathbf{A}}}$ is convenient, since the $\hat{\mathbf{J}}$ matrix will already have been computed and factored for back-solution if the Newton-Raphson iterating method is used. (In fact, Kent et al. [6] stop here in their error analysis.) However, the Gauss-Markov theorem assumes that the data are homoscedastic and that the independent variables are non-stochastic. In the WLS problem when the weights are scalars, a heteroscedastic data set can be converted to homoscedastic by scaling the data, but this cannot be done if the weights are non-diagonal matrices ($n_y > 1$). Besides that, the problem being addressed here assumes data with stochastic dependent variables. Therefore, as indicated, (37) holds only approximately in this case.

A more accurate estimate of $\mathbf{\Sigma}_{\hat{\mathbf{A}}}$ can be found by propagating the covariances of the data through the functional dependence of $\hat{\mathbf{A}}$ on the data points. On very mild assumptions about the distribution of the data errors, one obtains the well-known quadrature formula (see, e.g. Bevington [2]), which is accurate to first order in the variances:

$$\mathbf{\Sigma}_{\hat{\mathbf{A}}} \;\;=\;\; \sum_i \frac{\mathrm{d}\hat{\mathbf{A}}}{\mathrm{d}\varkappa_i} \mathbb{\Sigma}_i \left(\frac{\mathrm{d}\hat{\mathbf{A}}}{\mathrm{d}\varkappa_i} \right)^{\mathrm{T}}. \tag{38}$$

Since $\hat{\mathbf{A}}$ is implicitly defined in terms of the $\varkappa_i$ via $\hat{\mathbf{F}}(\hat{\mathbf{A}}) = \mathbf{0}$ with $\hat{\mathbf{F}}$ as in (30), $\mathrm{d}\hat{\mathbf{A}}/\mathrm{d}\varkappa_i$ is obtained by implicit differentiation and use of the chain rule:

$$\frac{\mathrm{d}\hat{\mathbf{F}}}{\mathrm{d}\varkappa_i} = \frac{\partial\hat{\mathbf{F}}}{\partial\varkappa_i} + \frac{\partial\hat{\mathbf{F}}}{\partial\hat{\mathbf{A}}} \frac{\mathrm{d}\hat{\mathbf{A}}}{\mathrm{d}\varkappa_i} = \frac{\partial\hat{\mathbf{F}}}{\partial\varkappa_i} + \hat{\mathbf{J}} \frac{\mathrm{d}\hat{\mathbf{A}}}{\mathrm{d}\varkappa_i} = \mathbf{0} \quad \Rightarrow \quad \frac{\mathrm{d}\hat{\mathbf{A}}}{\mathrm{d}\varkappa_i} = -\hat{\mathbf{J}}^{-1} \frac{\partial\hat{\mathbf{F}}}{\partial\varkappa_i}. \tag{39}$$

This gives

$$\mathbf{\Sigma}_{\hat{\mathbf{A}}} \;\;=\;\; \hat{\mathbf{J}}^{-1} \hat{\mathbf{Q}} \hat{\mathbf{J}}^{-1} \quad \text{where} \quad \hat{\mathbf{Q}} \doteq \sum_i \frac{\partial\hat{\mathbf{F}}}{\partial\varkappa_i} \mathbb{\Sigma}_i \left(\frac{\partial\hat{\mathbf{F}}}{\partial\varkappa_i} \right)^{\mathrm{T}}. \tag{40}$$

In practice, $\hat{\mathbf{J}}^{-1}$ does not need to be computed; instead (40) is written as $\hat{\mathbf{J}}\mathbf{\Sigma}_{\hat{\mathbf{A}}}\hat{\mathbf{J}} = \hat{\mathbf{Q}}$ which is solved by factoring $\hat{\mathbf{J}}$ and performing two back-solutions. The calculation of $\hat{\mathbf{Q}}$ is given in

the Appendix, Sec. 8.2.

$$\hat{\mathbf{Q}} \;=\; \sum_i \left\{ \hat{\boldsymbol{\mu}}_i \hat{\boldsymbol{\mu}}_i^{\mathrm{T}} \otimes \mathbf{W}_i + \boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_i} \otimes \mathbf{W}_i \boldsymbol{r}_i \boldsymbol{r}_i^{\mathrm{T}} \mathbf{W}_i - \hat{\boldsymbol{\Xi}}_i^{\;\mathrm{T}} \mathbf{W}_i \hat{\boldsymbol{\Xi}}_i \otimes \mathbf{W}_i \boldsymbol{r}_i \boldsymbol{r}_i^{\mathrm{T}} \mathbf{W}_i \right\}. \quad (41)$$

Conveniently, each of the terms in $\hat{\mathbf{Q}}$ also appears in $\hat{\mathbf{J}}$, so little additional work needs to be done to compute it. Comparing (41) with (31), it is seen that $\hat{\mathbf{Q}}$ has the same leading term as $\hat{\mathbf{J}}$, but that it lacks the two column-wise outer product terms present in $\hat{\mathbf{J}}$, and differs in sign with respect to the other two terms. The terms which differ are of first or second order in the residuals, so that by neglecting the residuals one obtains $\hat{\mathbf{J}}^{-1} \hat{\mathbf{Q}} \hat{\mathbf{J}}^{-1} \approx \hat{\mathbf{J}}^{-1}$, recovering (37) and showing that it is only an approximation.

Note that by (40), $\hat{\mathbf{Q}}$ is always positive definite, and therefore so is $\boldsymbol{\Sigma}_{\hat{\mathbf{A}}}$, as required for a covariance matrix.

If the value of χ^2_ν calculated using (36) departs significantly from unity, and there is reason to believe that this is due to systematic over- or underestimation of the uncertainties of the data rather than a failure of the assumed linear model, then the covariance matrix $\boldsymbol{\Sigma}_{\hat{\mathbf{A}}}$ can be multiplied by the calculated χ^2_ν, which is equivalent to multiplying all the data covariances by this factor. The same procedure should be used to adjust $\boldsymbol{\Sigma}_{\hat{\mathbf{A}}}$ if the $\boldsymbol{\Sigma}_i$ of the data points are only known up to a global multiplicative constant, e.g. when the data are assumed to have equal but unknown uncertainties in all dimensions.

4 Reduced-size system for unconstrained case

When solving the unconstrained problem, one can take advantage of the fact that the constant term $\boldsymbol{a}_0$ can be solved for in terms of the rest of $\hat{\mathbf{A}}$. Solving for it separately reduces the size of the linear system to be solved on each iteration. There is also often a significant reduction in the ill-conditioning of the linear system because the recasting of the equations is roughly (or for $n_y = 1$, precisely) equivalent to shifting the origin of coordinates to the centroid of the data. For example, in the eight-dimensional fit in the example of Sec. 5, the size of the Jacobian matrix is reduced from 18×18 to 12×12 and its 2-norm condition is reduced by a factor of ~ 30.

4.1 Reducing the size of the quasi-normal equations

In terms of "un-augmented" quantities, $\boldsymbol{r}_i = \boldsymbol{a}_0 + \mathbf{A}\boldsymbol{x}_i - \boldsymbol{y}_i$ and (27) becomes the system

$$\sum_i \mathbf{W}_i(\boldsymbol{a}_0 + \mathbf{A}\boldsymbol{x}_i) \;=\; \sum_i \mathbf{W}_i \boldsymbol{y}_i \tag{42}$$

$$\sum_i \mathbf{W}_i(\boldsymbol{a}_0 + \mathbf{A}\boldsymbol{x}_i)\boldsymbol{\mu}_i^{\mathrm{T}} \;=\; \sum_i \mathbf{W}_i \boldsymbol{y}_i \boldsymbol{\mu}_i^{\mathrm{T}}. \tag{43}$$

From (42)

$$\boldsymbol{a}_0 \;=\; -\boldsymbol{\Omega}^{-1} \sum_i \mathbf{W}_i(\mathbf{A}\boldsymbol{x}_i - \boldsymbol{y}_i) \tag{44}$$

where $\boldsymbol{\Omega} \doteq \sum_i \mathbf{W}_i$. Inserting this into (43) yields the reduced-size system of quasi-normal equations

$$\sum_i \mathbf{W}_i \left(\mathbf{A}\boldsymbol{x}_i - \boldsymbol{\Omega}^{-1} \sum_j \mathbf{W}_j \mathbf{A}\boldsymbol{x}_j \right) \boldsymbol{\mu}_i^{\mathrm{T}} =$$

$$\sum_i \mathbf{W}_i \left(\boldsymbol{y}_i - \boldsymbol{\Omega}^{-1} \sum_j \mathbf{W}_j \boldsymbol{y}_j \right) \boldsymbol{\mu}_i^{\mathrm{T}} \tag{45}$$

which can be solved by either the fixed-point or Newton-Raphson iteration methods. For the fixed-point method the basis is again to vectorize to remove vec $\mathbf{A}$ from the sums and solve iteratively, re-determining $\boldsymbol{a}_0$ from (44) after each update of $\mathbf{A}$. The reduction in system size is from $(n_x + 1)n_y$ to $n_x n_y$ linear equations, where $n_x = n_{\hat{x}} - 1$ is the size of the (un-augmented) $\boldsymbol{x}_i$, i.e. the number of independent coordinates.

4.2 Reduced-size form of Newton-Raphson iteration

The Jacobian matrix for Newton-Raphson iteration in the reduced-size formulation is more complicated to derive due to the extra terms in (45) as compared to (27), but the result turns out to be not much more complicated than (31). The derivation is given in the Appendix, Sec. 8.3. The result is

$$\begin{aligned}
\mathbf{J} &= \sum_i \Big\{ \boldsymbol{\mu}_i \boldsymbol{\mu}_i^{\mathrm{T}} \otimes \mathbf{W}_i - \boldsymbol{\Sigma}_{\boldsymbol{x}_i} \otimes \mathbf{W}_i \boldsymbol{r}_i \boldsymbol{r}_i^{\mathrm{T}} \mathbf{W}_i + \boldsymbol{\Xi}_i^{\mathrm{T}} \mathbf{W}_i \boldsymbol{\Xi}_i \otimes \mathbf{W}_i \boldsymbol{r}_i \boldsymbol{r}_i^{\mathrm{T}} \mathbf{W}_i \\
&\quad - \mathbf{W}_i \boldsymbol{\Xi}_i \bowtie \mathbf{W}_i \boldsymbol{r}_i \boldsymbol{\mu}_i^{\mathrm{T}} - \mathbf{W}_i \boldsymbol{r}_i \boldsymbol{\mu}_i^{\mathrm{T}} \bowtie \mathbf{W}_i \boldsymbol{\Xi}_i \Big\} - \langle \mathbf{Z} \rangle \boldsymbol{\Omega} \langle \mathbf{Z} \rangle^{\mathrm{T}}
\end{aligned} \tag{46}$$

where $\boldsymbol{\Xi}_i \doteq \mathbf{A} \boldsymbol{\Sigma}_{\boldsymbol{x}_i} - \boldsymbol{\Sigma}_{\boldsymbol{x}_i \boldsymbol{y}_i}$ and $\langle \mathbf{Z} \rangle \doteq \langle \mathbf{M} \rangle - \langle \mathbf{E} \rangle$, with

$$\langle \mathbf{M} \rangle \doteq \left(\sum_i \boldsymbol{\mu}_i \otimes \mathbf{W}_i \right) \boldsymbol{\Omega}^{-1}, \qquad \langle \mathbf{E} \rangle \doteq \left(\sum_i \boldsymbol{\Xi}_i^{\mathrm{T}} \mathbf{W}_i \otimes \mathbf{W}_i \boldsymbol{r}_i \right) \boldsymbol{\Omega}^{-1}. \tag{47}$$

Equation (46) is the same as (31) except for the removal of hats from all quantities and the additional term involving $\langle \mathbf{Z} \rangle$.

4.3 Reduced-size form of covariance of A

The covariance of the un-augmented $\mathbf{A}$ is $\boldsymbol{\Sigma}_{\mathbf{A}} = \mathbf{J}^{-1} \mathbf{Q} \mathbf{J}^{-1}$ with $\mathbf{J}$ defined in (46). The following expression for $\mathbf{Q}$ is worked out in the Appendix, Sec. 8.4:

$$\begin{aligned}
\mathbf{Q} &= \sum_i \Big\{ \boldsymbol{\mu}_i \boldsymbol{\mu}_i^{\mathrm{T}} \otimes \mathbf{W}_i + \boldsymbol{\Sigma}_{\boldsymbol{x}_i} \otimes \mathbf{W}_i \boldsymbol{r}_i \boldsymbol{r}_i^{\mathrm{T}} \mathbf{W}_i - \boldsymbol{\Xi}_i^{\mathrm{T}} \mathbf{W}_i \boldsymbol{\Xi}_i \otimes \mathbf{W}_i \boldsymbol{r}_i \boldsymbol{r}_i^{\mathrm{T}} \mathbf{W}_i \Big\} \\
&\quad - \langle \mathbf{M} \rangle \boldsymbol{\Omega} \langle \mathbf{M} \rangle^{\mathrm{T}} + \langle \mathbf{E} \rangle \boldsymbol{\Omega} \langle \mathbf{E} \rangle^{\mathrm{T}}.
\end{aligned} \tag{48}$$

In the reduced-size analysis, the covariance of $\boldsymbol{a}_0$ (the $n_y \times n_y$ upper left block of $\boldsymbol{\Sigma}_{\hat{\mathbf{A}}}$) must be calculated separately. It works out to

$$\boldsymbol{\Sigma}_{\boldsymbol{a}_0} = \boldsymbol{\Omega}^{-1} + \langle \mathbf{Z} \rangle^{\mathrm{T}} \boldsymbol{\Sigma}_{\mathbf{A}} \langle \mathbf{Z} \rangle - \langle \mathbf{E} \rangle^{\mathrm{T}} \mathbf{J}^{-1} \langle \mathbf{Z} \rangle - \langle \mathbf{Z} \rangle^{\mathrm{T}} \mathbf{J}^{-1} \langle \mathbf{E} \rangle. \tag{49}$$

The covariance of $\boldsymbol{a}_0$ with $\mathbf{A}$ (the $n_x n_y \times n_y$ lower left block of $\boldsymbol{\Sigma}_{\hat{\mathbf{A}}}$ and transpose of its upper right block) is

$$\boldsymbol{\Sigma}_{\boldsymbol{a}_0, \mathbf{A}} = \mathbf{J}^{-1} \langle \mathbf{E} \rangle - \boldsymbol{\Sigma}_{\mathbf{A}} \langle \mathbf{Z} \rangle. \tag{50}$$

4.4 Reduced-size form for the case of one dependent coordinate

When there is only one dependent coordinate ($n_y = 1$) some significant simplifications in the foregoing expressions are possible. This is a very common case and the simplifications

improve the numerical properties of the linear systems being solved, so the results are worth stating.

In (45) it becomes possible to move $\mathbf{A}$ (which is now a row vector $\boldsymbol{a}^{\mathrm{T}}$) out of the sums without resorting to the Kronecker product because $\mathbf{W}_i = W_i$ and $\boldsymbol{\Omega} = \Omega$ are scalars. One can then define the weighted mean $\langle \boldsymbol{x} \rangle = \Omega^{-1} \sum_j W_j \boldsymbol{x}_j$ and deviations from the mean $\boldsymbol{x}_i' = \boldsymbol{x}_i - \langle \boldsymbol{x} \rangle$, and in the same manner define $\boldsymbol{\mu}_i'$ and y_i' (the latter now scalar). Then (45) can be written

$$\boldsymbol{a}^{\mathrm{T}} \sum_i W_i \boldsymbol{x}_i'(\boldsymbol{\mu}_i')^{\mathrm{T}} = \sum_i W_i y_i'(\boldsymbol{\mu}_i')^{\mathrm{T}}. \tag{51}$$

For $n_x = 1$ (straight line in 2-D) using (51) for the fixed-point iteration is essentially Williamson's method [10].

Defining $\boldsymbol{\epsilon}_i = \boldsymbol{x}_i - \boldsymbol{\mu}_i$, and $\boldsymbol{\epsilon}_i' = \boldsymbol{\epsilon}_i - \langle \boldsymbol{\epsilon} \rangle$, (46) reduces to

$$\mathbf{J} = \sum_i W_i \left\{ (\boldsymbol{x}_i' - 2\boldsymbol{\epsilon}_i')(\boldsymbol{x}_i' - 2\boldsymbol{\epsilon}_i')^{\mathrm{T}} - W_i r_i^2 \boldsymbol{\Sigma}_{\boldsymbol{x}_i} \right\}. \tag{52}$$

The covariance expressions (48), (49) and (50) become respectively

$$\mathbf{Q} = \sum_i W_i \left\{ \boldsymbol{\mu}_i'(\boldsymbol{\mu}_i')^{\mathrm{T}} - \boldsymbol{\epsilon}_i'(\boldsymbol{\epsilon}_i')^{\mathrm{T}} + W_i r_i^2 \boldsymbol{\Sigma}_{\boldsymbol{x}_i} \right\} \tag{53}$$

$$\boldsymbol{\Sigma}_{a_0} = \Omega^{-1} + (\langle \boldsymbol{x} \rangle - 2\langle \boldsymbol{\epsilon} \rangle)^{\mathrm{T}} \boldsymbol{\Sigma}_{\mathbf{a}} (\langle \boldsymbol{x} \rangle - 2\langle \boldsymbol{\epsilon} \rangle) - 2 (\langle \boldsymbol{x} \rangle - 2\langle \boldsymbol{\epsilon} \rangle)^{\mathrm{T}} \mathbf{J}^{-1} \langle \boldsymbol{\epsilon} \rangle \tag{54}$$

$$\boldsymbol{\Sigma}_{a_0,\mathbf{a}} = \mathbf{J}^{-1} \langle \boldsymbol{\epsilon} \rangle - \boldsymbol{\Sigma}_{\mathbf{a}} (\langle \boldsymbol{x} \rangle - 2\langle \boldsymbol{\epsilon} \rangle). \tag{55}$$

5 Application: xenon isotopes

As an example, this method is applied to a re-analysis of the data in a classic paper by Reynolds et al. [9] on the study of xenon isotopes from the Apollo 14 lunar breccia 14318. I selected this paper because it introduced what was at the time a novel method of solving, with a minimum of assumptions, a recurring problem in that field, namely identifying the isotopic composition of an unknown component of a mixture. Reynolds et al. used their method to demonstrate the presence of ^{244}Pu (half-life 81 Ma) in the early Moon. Their method of analysis has been widely adopted by later workers; however, this method, while adequate to establish the conclusion of their paper, was somewhat defective from a statistical standpoint, as will be described later. The method developed here, while following the same basic plan, remedies these defects.

The measured isotopic compositions are expected to be linear combinations of contributions from three sources: trapped gas, with an isotopic composition close to the solar value; cosmic-ray spallation of elements near Xe in atomic mass, which produces all isotopes in roughly comparable amounts; and spontaneous fission of heavy elements, which only contributes to the isotopes reachable by a beta-decay chain from the neutron-rich side.

Consider first the simplified situation in which only three isotopic ratios are used. Suppose a 3-dimensional plot is constructed with $x = {}^{126}\mathrm{Xe}/{}^{132}\mathrm{Xe}$, $y = {}^{130}\mathrm{Xe}/{}^{132}\mathrm{Xe}$, and $z = {}^{136}\mathrm{Xe}/{}^{132}\mathrm{Xe}$. If each measured sample is a mixture (a linear combination) of three distinct components, then it must lie on a plane within this space. More specifically, it must lie within the triangle whose vertices are the three component compositions. Therefore fitting a plane to the data provides a constraint on these compositions. (Since the data were measured directly as isotopic ratios and not as abundances, the plane is defined in terms of ratios and is not constrained to pass through the origin.) For this plane to be adequately defined by the data, it is necessary for different measurements to include significantly different

proportions of the three components. Reynolds et al. used the method of stepwise thermal release, in which the gas released at each of a series of successively higher temperatures is measured. This works because the different components are sited in locations that have different thermal release patterns. In order to determine the fissiogenic component, one can make use of the fact that the beta-shielded isotopes ^{126}Xe and ^{130}Xe are not present in it, so the fissiogenic ratio ^{136}Xe/^{132}Xe is found by the intercept of the fitted plane with the z-axis.

Reynolds et al. repeated this process using other isotopes in turn in place of ^{136}Xe to determine the complete fissiogenic spectrum, which they then re-normalized to ^{136}Xe $\equiv 1$. However, this procedure of performing fits to selected sets of three isotopic ratios at a time does not guarantee that the fits are compatible with one another, because there is no requirement that the Deming adjusted values for the independent variables (^{126}Xe/^{132}Xe and ^{130}Xe/^{132}Xe) will be the same for all the fits. This difficulty can be avoided by performing a single unified fit with these two isotopes as independent variables and all the other isotopes as dependent variables. Since xenon has 9 stable isotopes, there are 8 isotopic ratios with respect to any chosen reference isotope, so the fit including all isotopes will have 2 independent coordinates and 6 dependent coordinates.

A small complication is introduced by the fact that the mass spectrometric measurements determined the ratios of the various isotopes with respect to ^{130}Xe rather than to ^{132}Xe or another fissiogenic isotope. The data still should lie on a plane, but since ^{130}Xe is a beta-shielded isotope, the fissiogenic component lies at infinity in this space. Reynolds et al. chose to re-normalize the data to ^{132}Xe so that the z-intercept could be used as described above to obtain the fissiogenic production ratios. Doing this, however, introduces covariances among the data that were not included in the fitting method they used. But there is a simple way to avoid the need for changing the reference isotope. Returning to the simplified example, we use the same four isotopes but perform a fit of the form:

$$^{132}\text{Xe}/^{130}\text{Xe} = a_0 + a_1(^{126}\text{Xe}/^{130}\text{Xe}) + a_2(^{136}\text{Xe}/^{130}\text{Xe}). \tag{56}$$

Now multiply this equation by ^{130}Xe/^{136}Xe so that ^{136}Xe becomes the reference isotope:

$$^{132}\text{Xe}/^{136}\text{Xe} = a_0(^{130}\text{Xe}/^{136}\text{Xe}) + a_1(^{126}\text{Xe}/^{136}\text{Xe}) + a_2. \tag{57}$$

Setting the two beta-shielded isotopic ratios to zero yields a_2 as the fissiogenic value of ^{132}Xe/^{136}Xe. For the multi-dimensional fit, one obtains the $\hat{\mathbf{A}}$ matrix with multiple rows, one for each of the dependent variables, so the whole fissiogenic composition (except for ^{126}Xe/^{136}Xe and ^{130}Xe/^{136}Xe, which are zero by hypothesis) is obtained as the last column of this matrix. The variances and covariances of these values are obtained from the corresponding entries of $\mathbf{\Sigma}_{\hat{\mathbf{A}}}$.

The results of this eight-dimensional (8D) fit are shown in Fig. 1. In addition to the measurements of breccia 14318, the data set used for the fits includes a probable trapped composition, known as SUCOR, derived as the surface-correlated component measured in lunar fines. Including this component helps anchor the fitted plane and its beneficial effect is confirmed by observing that the condition of the quasi-normal equations is reduced by a factor of ~ 3 (to 5.2×10^4 in the reduced-size formulation) while χ^2_ν remains about the same. The results obtained by Reynolds et al. using their method on the same data set are also shown for comparison. As can be seen from the figure, the present method gives results that agree within uncertainties with those of Reynolds et al., despite the differences of method. The computed uncertainties in the coefficients are lower, probably through avoiding the changes of reference isotope.

The 8D fit also includes two isotopes not included in the fits done by Reynolds et al. The value for ^{124}Xe, which is consistent with zero as expected for a beta-shielded isotope,

Method	Apparent fission spectrum						
	^{124}Xe	^{128}Xe	^{129}Xe	^{131}Xe	^{132}Xe	^{134}Xe	^{136}Xe
This work	−0.00092 ±0.00076	0.0194 ±0.0041	0.027 ±0.024	0.313 ±0.016	0.871 ±0.022	0.904 ±0.013	≡ 1
Ref. [9]	—	—	0.031 ±0.043	0.297 ±0.031	0.870 ±0.030	0.903 ±0.043	≡ 1

Figure 1: Results of least-squares fits to the data of Reynolds et al. [9].

provides an internal check on the method. The value for ^{128}Xe, another beta-shielded isotope, is positive by more than 4σ. This nonzero value probably arises from a contribution to this isotope from capture of cosmic-ray secondary neutrons on ^{127}I, which evidently correlates with the fission component.

For this fit, $\chi^2_\nu = 1.90$. This is significantly greater than unity, indicating that the three-component model is not sufficient to account for all of the variation of the data. A detailed examination shows that ^{131}Xe alone contributes about half of the total χ^2. Probably this scatter is due to a contribution from neutron-capture on ^{130}Te. This explanation is consistent with the probable neutron-capture contribution to ^{128}Xe, but for ^{131}Xe this component does not seem to be correlated with the fission component. If it is correlated to some extent, it can increase the value of the apparent fission ratio of ^{131}Xe/^{136}Xe, but it is not possible with the data on hand to assess by how much. The most one can say is that the value determined by the fit is an upper limit to the true fission ratio for this isotope.

In meteorites, the isotope ^{129}Xe can contain a contribution from beta decay of now-extinct iodine ^{129}I (half-life 16 Ma), but here ^{129}Xe contributes only about 10% of the total χ^2, so this component can be at most a minor contribution in this lunar breccia.

Repeating the fit excluding the isotopes ^{128}Xe, ^{129}Xe, and ^{131}Xe to eliminate the neutron-capture and possible ^{129}I-derived components gives virtually the same apparent fission yields for ^{124}Xe, ^{132}Xe and ^{134}Xe as the 8D fit, but with $\chi^2_\nu = 0.90$, now quite close to unity. This result supports the hypothesis that the six included isotopes form a 3-component system, and that the quoted uncertainties of the data points are not substantially over- or underestimated.

6 Conclusion

The method presented here fits a general linear (affine) relationship between the independent and dependent variables to data with variances and covariances that are known *a priori,* at least up to a single overall multiplicative constant. For the fit not constrained to pass through the origin, solving separately for the constant term is recommended because it reduces both the size and, commonly, the condition of the linear system to be solved on each iteration. The method can be used for the ordinary unweighted and weighted least-squares problems (by suitably defining the covariance matrices), although more efficient techniques for solving those problems exist. It also encompasses the methods of Williamson [10] and of Kent et al. [6] as special cases in which there is only one dependent variable.

Kent et al. show how groups of related data that share one or more of the parameters of the fit can be accommodated in a single unified treatment. The method works by defining new coordinates for the unrelated portions of each group, artificially expanding the data vector x. For each data point x_i, the coordinates corresponding to the groups to which it does not belong are set to 0, and their variances and covariances are set to 0 in Σ_i. This technique is equally usable with the more general method presented here.

inary development of this method.

7 Appendix: matrix direct products

7.1 Kronecker Product

If $\mathbf{A}$ is $m \times p$ and $\mathbf{B}$ is $q \times n$, then the Kronecker product of $\mathbf{A}$ and $\mathbf{B}$ is defined as

$$\mathbf{A} \otimes \mathbf{B} \;\doteq\; \left[\begin{array}{ccc} a_{11}\mathbf{B} & \cdots & a_{1p}\mathbf{B} \\ \vdots & \vdots & \vdots \\ a_{m1}\mathbf{B} & \cdots & a_{mp}\mathbf{B} \end{array} \right] \tag{58}$$

This has mq rows and pn columns. We assign it a precedence higher than matrix addition/subtraction, but lower than matrix multiplication. The usefulness of the Kronecker product for us is that if $\mathbf{X}$ is $p \times q$ so that $\mathbf{A}\,\mathbf{X}\,\mathbf{B}$ is defined, then

$$\mathrm{vec}\,(\mathbf{A}\,\mathbf{X}\,\mathbf{B}) \;\;=\;\; (\mathbf{B}^{\mathrm{T}} \otimes \mathbf{A})\,\mathrm{vec}\,\mathbf{X}. \tag{59}$$

Properties include (provided the indicated matrix operations are defined):

$(\mathbf{A} \otimes \mathbf{B}) \otimes \mathbf{C} = \mathbf{A} \otimes (\mathbf{B} \otimes \mathbf{C})$

$\mathbf{A} \otimes (\mathbf{B} + \mathbf{C}) = \mathbf{A} \otimes \mathbf{B} + \mathbf{A} \otimes \mathbf{C}$

$(\mathbf{A} \otimes \mathbf{B})^{\mathrm{T}} = \mathbf{A}^{\mathrm{T}} \otimes \mathbf{B}^{\mathrm{T}}$

$\boldsymbol{a} \otimes (\mathbf{B}\mathbf{C}) = (\boldsymbol{a} \otimes \mathbf{B})\mathbf{C}$

$\boldsymbol{a}^{\mathrm{T}} \otimes (\mathbf{B}\mathbf{C}) = \mathbf{B}(\boldsymbol{a}^{\mathrm{T}} \otimes \mathbf{C})$

$(\mathbf{B}\mathbf{C}) \otimes \boldsymbol{a} = (\mathbf{B} \otimes \boldsymbol{a})\mathbf{C}$

$(\mathbf{B}\mathbf{C}) \otimes \boldsymbol{a}^{\mathrm{T}} = \mathbf{B}(\mathbf{C} \otimes \boldsymbol{a}^{\mathrm{T}})$

$(\mathbf{A} \otimes \mathbf{B})^{\mathrm{T}} = \mathbf{A}^{\mathrm{T}} \otimes \mathbf{B}^{\mathrm{T}}$

$(\mathbf{A} \otimes \mathbf{B})^{-1} = \mathbf{A}^{-1} \otimes \mathbf{B}^{-1}$

$(\mathbf{A}\mathbf{B}) \otimes (\mathbf{C}\mathbf{D}) = (\mathbf{A} \otimes \mathbf{C})(\mathbf{B} \otimes \mathbf{D})$

$\boldsymbol{a} \otimes \boldsymbol{b} = \mathrm{vec}\,(\boldsymbol{b}\boldsymbol{a}^{\mathrm{T}})$

$\boldsymbol{a} \otimes \boldsymbol{b}^{\mathrm{T}} = \boldsymbol{a}\boldsymbol{b}^{\mathrm{T}}$

See [8] for more details.

7.2 Columnwise Outer Product

Now suppose that $\mathbf{X}$ is again $p \times q$ while now $\mathbf{A}$ is $m \times q$ and $\mathbf{B}$ is $p \times n$, so that $\mathbf{A}\,\mathbf{X}^{\mathrm{T}}\mathbf{B}$ is defined, then

$$\begin{aligned} \mathrm{vec}\,(\mathbf{A}\,\mathbf{X}^{\mathrm{T}}\mathbf{B}) \;\;&=\;\; (\mathbf{B}^{\mathrm{T}} \otimes \mathbf{A})\,\mathrm{vec}\,\mathbf{X}^{\mathrm{T}} \\ &=\;\; (\mathbf{B}^{\mathrm{T}} \otimes \mathbf{A})\,\mathbf{K}_{p,q}\mathrm{vec}\,\mathbf{X} \end{aligned} \tag{60}$$

where $\mathbf{K}_{p,q}$ is the permutation matrix (called the commutation matrix by Magnus and Neudecker [8]) that turns $\operatorname{vec}\mathbf{X}$ into $\operatorname{vec}\mathbf{X}^{\mathrm{T}}$. I choose to avoid using the commutation matrix and instead define a new matrix direct product,

$$\mathbf{A} \bowtie \mathbf{B} \;\;\doteq\;\; (\mathbf{B}^{\mathrm{T}} \otimes \mathbf{A})\,\mathbf{K}_{p,q} \tag{61}$$

so that

$$\operatorname{vec}(\mathbf{A}\,\mathbf{X}^{\mathrm{T}}\mathbf{B}) \;\;=\;\; (\mathbf{A} \bowtie \mathbf{B})\operatorname{vec}\mathbf{X}^{\mathrm{T}}. \tag{62}$$

The reason for introducing this operator is that it has its own simple definition:

$$\mathbf{A} \bowtie \mathbf{B} \;\;\doteq\;\; \begin{bmatrix} \boldsymbol{a}_1\boldsymbol{b}_1^{\mathrm{T}} & \cdots & \boldsymbol{a}_p\boldsymbol{b}_1^{\mathrm{T}} \\ \vdots & \vdots & \vdots \\ \boldsymbol{a}_1\boldsymbol{b}_n^{\mathrm{T}} & \cdots & \boldsymbol{a}_p\boldsymbol{b}_n^{\mathrm{T}} \end{bmatrix}. \tag{63}$$

This has mn rows and pq columns. We assign it a precedence higher than matrix addition/subtraction, but lower than matrix multiplication. This can be called the columnwise outer product of $\mathbf{A}$ and $\mathbf{B}$.

The result in (63) does not seem to have been reported previously. It can be seen readily by construction, but Magnus [7] has kindly provided an elegant proof: write

$$\mathbf{B} = \sum_i \boldsymbol{b}_i\boldsymbol{e}_i^{\mathrm{T}} \quad \text{and} \quad \mathbf{A} = \sum_j \boldsymbol{a}_j\boldsymbol{f}_j^{\mathrm{T}} \tag{64}$$

where $\boldsymbol{e}_i$ and $\boldsymbol{f}_j$ are appropriately-sized unit vectors. Then

$$\mathbf{B}^{\mathrm{T}} \otimes \mathbf{A} \;\;=\;\; \sum_{ij}(\boldsymbol{e}_i\boldsymbol{b}_i^{\mathrm{T}}) \otimes (\boldsymbol{a}_j\boldsymbol{f}_j^{\mathrm{T}}) = \sum_{ij}(\boldsymbol{e}_i \otimes \boldsymbol{a}_j)(\boldsymbol{b}_i \otimes \boldsymbol{f}_j)^{\mathrm{T}}. \tag{65}$$

Now, using the properties [8] that $\mathbf{K}_{p,q}^{\mathrm{T}} = \mathbf{K}_{q,p}$ and $\mathbf{K}_{q,p}(\boldsymbol{x} \otimes \boldsymbol{y}) = \boldsymbol{y} \otimes \boldsymbol{x}$ for any p-vector $\boldsymbol{x}$ and q-vector $\boldsymbol{y}$,

$$(\mathbf{B}^{\mathrm{T}} \otimes \mathbf{A})\mathbf{K}_{p,q} \;\;=\;\; \sum_{ij}(\boldsymbol{e}_i \otimes \boldsymbol{a}_j)(\boldsymbol{f}_j \otimes \boldsymbol{b}_i)^{\mathrm{T}} = \sum_{ij}(\boldsymbol{e}_i\boldsymbol{f}_j^{\mathrm{T}}) \otimes (\boldsymbol{a}_j\boldsymbol{b}_i^{\mathrm{T}}) \tag{66}$$

which is the desired result.

The following properties of the columnwise outer product are easily proved (assuming the indicated matrix operations are defined):

$\mathbf{A} \bowtie (\mathbf{B} + \mathbf{C}) = \mathbf{A} \bowtie \mathbf{B} + \mathbf{A} \bowtie \mathbf{C}$

$(\mathbf{A} + \mathbf{B}) \bowtie \mathbf{C} = \mathbf{A} \bowtie \mathbf{C} + \mathbf{B} \bowtie \mathbf{C}$

$(\mathbf{A} \bowtie \mathbf{B})^{\mathrm{T}} = \mathbf{B} \bowtie \mathbf{A}$

$\mathbf{A}(\mathbf{B} \bowtie \boldsymbol{c}) = (\mathbf{AB}) \bowtie \boldsymbol{c}$

$(\boldsymbol{a} \bowtie \mathbf{B})\mathbf{C} = \boldsymbol{a} \bowtie (\mathbf{C}^{\mathrm{T}}\mathbf{B})$

$\boldsymbol{a} \bowtie \boldsymbol{b} = \boldsymbol{b}^{\mathrm{T}} \bowtie \boldsymbol{a}^{\mathrm{T}} = \boldsymbol{a}\boldsymbol{b}^{\mathrm{T}}$

7.3 Relationship Between Products

There are also some identities involving both $\otimes$ and $\bowtie$:

$$\mathbf{A} \otimes \boldsymbol{b} = \boldsymbol{b} \bowtie \mathbf{A}^{\mathrm{T}}$$

$$\mathbf{A} \otimes \boldsymbol{b}^{\mathrm{T}} = \mathbf{A} \bowtie \boldsymbol{b}$$

$$\boldsymbol{a} \otimes \mathbf{B} = \mathbf{B} \bowtie \boldsymbol{a}^{\mathrm{T}}$$

$$\boldsymbol{a}^{\mathrm{T}} \otimes \mathbf{B} = \boldsymbol{a}^{\mathrm{T}} \bowtie \mathbf{B}^{\mathrm{T}}$$

$$\boldsymbol{a} \otimes (\mathbf{B} \bowtie \boldsymbol{c}) = \mathbf{B} \bowtie \boldsymbol{c}^{\mathrm{T}} \boldsymbol{a}$$

$$\boldsymbol{a} \bowtie (\boldsymbol{b} \bowtie \mathbf{C}) = \mathbf{C} \otimes \boldsymbol{a}\boldsymbol{b}^{\mathrm{T}}$$

$$(\mathbf{A}\mathbf{B}) \otimes (\mathbf{C}\mathbf{D}) = (\mathbf{C} \bowtie \mathbf{A}^{\mathrm{T}})(\mathbf{B} \bowtie \mathbf{D}^{\mathrm{T}})$$

$$(\mathbf{A}\mathbf{B}) \bowtie (\mathbf{C}\mathbf{D}) = (\mathbf{D}^{\mathrm{T}} \otimes \mathbf{A})(\mathbf{B} \bowtie \mathbf{C})$$

$$(\mathbf{A}\mathbf{B}) \bowtie (\mathbf{C}\mathbf{D}) = (\mathbf{A} \bowtie \mathbf{D})(\mathbf{B} \otimes \mathbf{C}^{\mathrm{T}})$$

8 Appendix: Derivations

In this section we provide the detailed derivations of the results presented in the main body of the paper (Sec. 2.3, 3.2, 4.2, and 4.3).

8.1 Derivation of Jacobian matrix $\hat{\mathbf{J}}$

Taking the derivative of (30) and using the properties of the Kronecker product $\otimes$ and the column-wise outer product $\bowtie$ defined in the previous section,

$$\hat{\mathbf{J}} = \frac{\partial \hat{\mathbf{F}}}{\partial \hat{\mathbf{A}}} = \sum_i \left\{ \left(\hat{\boldsymbol{\mu}}_i \boldsymbol{r}_i^{\mathrm{T}} \otimes \mathbf{I}_{n_y}\right) \frac{\partial \mathbf{W}_i}{\partial \hat{\mathbf{A}}} + \left(\hat{\boldsymbol{\mu}}_i \otimes \mathbf{W}_i\right) \frac{\partial \mathbf{r}_i}{\partial \hat{\mathbf{A}}} + \left(\mathbf{W}_i \boldsymbol{r}_i \bowtie \mathbf{I}_{n_{\hat{x}}}\right) \frac{\partial \hat{\boldsymbol{\mu}}_i}{\partial \hat{\mathbf{A}}} \right\}$$

Now, from (19), (20) and (25),

$$
\begin{aligned}
\frac{\partial \mathbf{W}_i}{\partial \hat{\mathbf{A}}} &= -(\mathbf{W}_i \otimes \mathbf{W}_i) \frac{\partial \mathbf{W}_i^{-1}}{\partial \hat{\mathbf{A}}} \\
&= -(\mathbf{W}_i \otimes \mathbf{W}_i) \frac{\partial}{\partial \hat{\mathbf{A}}} \left(\hat{\mathbf{A}} \boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_i} \hat{\mathbf{A}}^{\mathrm{T}} - \hat{\mathbf{A}} \boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_i \boldsymbol{y}_i}^{\mathrm{T}} - \boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_i \boldsymbol{y}_i} \hat{\mathbf{A}}^{\mathrm{T}} + \boldsymbol{\Sigma}_{\boldsymbol{y}_i} \right) \\
&= -(\mathbf{W}_i \otimes \mathbf{W}_i) \left[(\hat{\mathbf{A}}\boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_i} \otimes \mathbf{I}_{n_y}) + (\hat{\mathbf{A}}\boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_i} \bowtie \mathbf{I}_{n_y}) - (\boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_i \boldsymbol{y}_i} \otimes \mathbf{I}_{n_y}) - (\boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_i \boldsymbol{y}_i} \bowtie \mathbf{I}_{n_y}) \right] \\
&= -\left[\mathbf{W}_i(\hat{\mathbf{A}}\boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_i} - \boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_i \boldsymbol{y}_i}) \otimes \mathbf{W}_i + \mathbf{W}_i(\hat{\mathbf{A}}\boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_i} - \boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_i \boldsymbol{y}_i}) \bowtie \mathbf{W}_i \right] \\
&= -\left[\mathbf{W}_i \hat{\boldsymbol{\Xi}}_i \otimes \mathbf{W}_i + \mathbf{W}_i \hat{\boldsymbol{\Xi}}_i \bowtie \mathbf{W}_i \right], \quad\quad\quad\quad (67) \\
\frac{\partial \mathbf{r}_i}{\partial \hat{\mathbf{A}}} &= \frac{\partial}{\partial \hat{\mathbf{A}}} \left(\hat{\mathbf{A}}\hat{\boldsymbol{x}}_i - \boldsymbol{y}_i \right) \\
&= \hat{\boldsymbol{x}}_i^{\mathrm{T}} \otimes \mathbf{I}_{n_y}, \quad\quad\quad\quad (68) \\
\frac{\partial \hat{\boldsymbol{\mu}}_i}{\partial \hat{\mathbf{A}}} &= \frac{\partial}{\partial \hat{\mathbf{A}}} \left[\hat{\boldsymbol{x}}_i - (\hat{\mathbf{A}}\boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_i} - \boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_i \boldsymbol{y}_i})^{\mathrm{T}} \mathbf{W}_i \boldsymbol{r}_i \right] \\
&= -\left[\boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_i} \bowtie \mathbf{W}_i \boldsymbol{r}_i + \left(\boldsymbol{r}_i^{\mathrm{T}} \otimes (\hat{\mathbf{A}}\boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_i} - \boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_i \boldsymbol{y}_i})^{\mathrm{T}} \right) \frac{\partial \mathbf{W}_i}{\partial \hat{\mathbf{A}}} \right.
\end{aligned}
$$

$$+ (\hat{\mathbf{A}}\mathbf{\Sigma}_{\hat{\boldsymbol{x}}_i} - \mathbf{\Sigma}_{\hat{\boldsymbol{x}}_i \boldsymbol{y}_i})^{\mathrm{T}} \mathbf{W}_i \frac{\partial \mathbf{r}_i}{\partial \hat{\mathbf{A}}} \Big]$$

$$= -\Big[\mathbf{\Sigma}_{\hat{\boldsymbol{x}}_i} \bowtie \mathbf{W}_i \boldsymbol{r}_i - \left(\boldsymbol{r}_i^{\mathrm{T}} \otimes \hat{\mathbf{\Xi}}_i^{\mathrm{T}} \right) \left(\mathbf{W}_i \hat{\mathbf{\Xi}}_i \otimes \mathbf{W}_i + \mathbf{W}_i \hat{\mathbf{\Xi}}_i \bowtie \mathbf{W}_i \right)$$

$$\qquad + \hat{\mathbf{\Xi}}_i^{\mathrm{T}} \mathbf{W}_i \left(\hat{\boldsymbol{x}}_i^{\mathrm{T}} \otimes \mathbf{I}_{n_y} \right) \Big]$$

$$= -\mathbf{\Sigma}_{\hat{\boldsymbol{x}}_i} \bowtie \mathbf{W}_i \boldsymbol{r}_i + \boldsymbol{r}_i^{\mathrm{T}} \mathbf{W}_i \hat{\mathbf{\Xi}}_i \otimes \hat{\mathbf{\Xi}}_i^{\mathrm{T}} \mathbf{W}_i + \hat{\mathbf{\Xi}}_i^{\mathrm{T}} \mathbf{W}_i \hat{\mathbf{\Xi}}_i \bowtie \mathbf{W}_i \boldsymbol{r}_i - \hat{\boldsymbol{x}}_i^{\mathrm{T}} \otimes \hat{\mathbf{\Xi}}_i^{\mathrm{T}} \mathbf{W}_i. \qquad (69)$$

So

$$\hat{\mathbf{J}} = \sum_i \Big\{ -\left(\hat{\boldsymbol{\mu}}_i \boldsymbol{r}_i^{\mathrm{T}} \otimes \mathbf{I}_{n_y} \right) \left[\mathbf{W}_i \hat{\mathbf{\Xi}}_i \otimes \mathbf{W}_i + \mathbf{W}_i \hat{\mathbf{\Xi}}_i \bowtie \mathbf{W}_i \right]$$

$$\qquad + \left(\hat{\boldsymbol{\mu}}_i \otimes \mathbf{W}_i \right) \left(\hat{\boldsymbol{x}}_i^{\mathrm{T}} \otimes \mathbf{I}_{n_y} \right)$$

$$\qquad + \left(\mathbf{W}_i \boldsymbol{r}_i \bowtie \mathbf{I}_{n_{\hat{x}}} \right) \left[-\mathbf{\Sigma}_{\hat{\boldsymbol{x}}_i} \bowtie \mathbf{W}_i \boldsymbol{r}_i \right.$$

$$\qquad \qquad + \boldsymbol{r}_i^{\mathrm{T}} \mathbf{W}_i \hat{\mathbf{\Xi}}_i \otimes \hat{\mathbf{\Xi}}_i^{\mathrm{T}} \mathbf{W}_i + \hat{\mathbf{\Xi}}_i^{\mathrm{T}} \mathbf{W}_i \hat{\mathbf{\Xi}}_i \bowtie \mathbf{W}_i \boldsymbol{r}_i$$

$$\qquad \qquad - \hat{\boldsymbol{x}}_i^{\mathrm{T}} \otimes \hat{\mathbf{\Xi}}_i^{\mathrm{T}} \mathbf{W}_i \big] \Big\} \qquad\qquad (70)$$

$$= \sum_i \Big\{ -\hat{\boldsymbol{\mu}}_i \boldsymbol{r}_i^{\mathrm{T}} \mathbf{W}_i \hat{\mathbf{\Xi}}_i \otimes \mathbf{W}_i - \mathbf{W}_i \hat{\mathbf{\Xi}}_i \bowtie \mathbf{W}_i \boldsymbol{r}_i \hat{\boldsymbol{\mu}}_i^{\mathrm{T}}$$

$$\qquad + \hat{\boldsymbol{\mu}}_i \hat{\boldsymbol{x}}_i^{\mathrm{T}} \otimes \mathbf{W}_i$$

$$\qquad - \mathbf{\Sigma}_{\hat{\boldsymbol{x}}_i} \otimes \mathbf{W}_i \boldsymbol{r}_i \boldsymbol{r}_i^{\mathrm{T}} \mathbf{W}_i$$

$$\qquad + \mathbf{W}_i \boldsymbol{r}_i \boldsymbol{r}_i^{\mathrm{T}} \mathbf{W}_i \hat{\mathbf{\Xi}}_i \bowtie \mathbf{W}_i \hat{\mathbf{\Xi}}_i + \hat{\mathbf{\Xi}}_i^{\mathrm{T}} \mathbf{W}_i \hat{\mathbf{\Xi}}_i \otimes \mathbf{W}_i \boldsymbol{r}_i \boldsymbol{r}_i^{\mathrm{T}} \mathbf{W}_i$$

$$\qquad - \mathbf{W}_i \boldsymbol{r}_i \hat{\boldsymbol{x}}_i^{\mathrm{T}} \bowtie \mathbf{W}_i \hat{\mathbf{\Xi}}_i \Big\}$$

Now put $\hat{\mathbf{\Xi}}_i^{\mathrm{T}} \mathbf{W}_i \boldsymbol{r}_i = \hat{\boldsymbol{\epsilon}}_i = \hat{\boldsymbol{x}}_i - \hat{\boldsymbol{\mu}}_i$ and collect terms.

$$\hat{\mathbf{J}} = \sum_i \Big\{ \hat{\boldsymbol{\mu}}_i \hat{\boldsymbol{\mu}}_i^{\mathrm{T}} \otimes \mathbf{W}_i - \mathbf{\Sigma}_{\hat{\boldsymbol{x}}_i} \otimes \mathbf{W}_i \boldsymbol{r}_i \boldsymbol{r}_i^{\mathrm{T}} \mathbf{W}_i + \hat{\mathbf{\Xi}}_i^{\mathrm{T}} \mathbf{W}_i \hat{\mathbf{\Xi}}_i \otimes \mathbf{W}_i \boldsymbol{r}_i \boldsymbol{r}_i^{\mathrm{T}} \mathbf{W}_i$$

$$\qquad - \mathbf{W}_i \hat{\mathbf{\Xi}}_i \bowtie \mathbf{W}_i \boldsymbol{r}_i \hat{\boldsymbol{\mu}}_i^{\mathrm{T}} - \mathbf{W}_i \boldsymbol{r}_i \hat{\boldsymbol{\mu}}_i^{\mathrm{T}} \bowtie \mathbf{W}_i \hat{\mathbf{\Xi}}_i \Big\} . \qquad\qquad (71)$$

8.2 Derivation of matrix $\hat{\mathbf{Q}}$

Here we derive the matrix $\hat{\mathbf{Q}}$ defined in (40) in Sec. 3.2. Define

$$\hat{\mathbf{F}}_{\hat{\boldsymbol{x}}_i} \doteq \frac{\partial \hat{\mathbf{F}}}{\partial \hat{\boldsymbol{x}}_i} = \hat{\boldsymbol{\mu}}_i \otimes \mathbf{W}_i \hat{\mathbf{A}} + \mathbf{W}_i \boldsymbol{r}_i \bowtie \left(\mathbf{I}_{n_{\hat{x}}} - \hat{\mathbf{A}}^{\mathrm{T}} \mathbf{W}_i \hat{\mathbf{\Xi}}_i \right)$$

$$\hat{\mathbf{F}}_{\boldsymbol{y}_i} \doteq \frac{\partial \hat{\mathbf{F}}}{\partial \boldsymbol{y}_i} = -\hat{\boldsymbol{\mu}}_i \otimes \mathbf{W}_i + \mathbf{W}_i \boldsymbol{r}_i \bowtie \mathbf{W}_i \hat{\mathbf{\Xi}}_i$$

Then

$$\frac{\partial \hat{\mathbf{F}}}{\partial \boldsymbol{z}_i} \mathbf{\Sigma}_i \left(\frac{\partial \hat{\mathbf{F}}}{\partial \boldsymbol{z}_i} \right)^{\mathrm{T}} = \left[\hat{\mathbf{F}}_{\hat{\boldsymbol{x}}_i}, \ \hat{\mathbf{F}}_{\boldsymbol{y}_i} \right] \begin{bmatrix} \mathbf{\Sigma}_{\hat{\boldsymbol{x}}_i} & \mathbf{\Sigma}_{\hat{\boldsymbol{x}}_i \boldsymbol{y}_i}^{\mathrm{T}} \\ \mathbf{\Sigma}_{\hat{\boldsymbol{x}}_i \boldsymbol{y}_i} & \mathbf{\Sigma}_{\boldsymbol{y}_i} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{F}}_{\hat{\boldsymbol{x}}_i}^{\mathrm{T}} \\ \hat{\mathbf{F}}_{\boldsymbol{y}_i}^{\mathrm{T}} \end{bmatrix}$$

$$= \hat{\mathbf{F}}_{\hat{\boldsymbol{x}}_i} \mathbf{\Sigma}_{\hat{\boldsymbol{x}}_i} \hat{\mathbf{F}}_{\hat{\boldsymbol{x}}_i}^{\mathrm{T}} + \hat{\mathbf{F}}_{\hat{\boldsymbol{x}}_i} \mathbf{\Sigma}_{\hat{\boldsymbol{x}}_i \boldsymbol{y}_i}^{\mathrm{T}} \hat{\mathbf{F}}_{\boldsymbol{y}_i}^{\mathrm{T}} + \hat{\mathbf{F}}_{\boldsymbol{y}_i} \mathbf{\Sigma}_{\hat{\boldsymbol{x}}_i \boldsymbol{y}_i} \hat{\mathbf{F}}_{\hat{\boldsymbol{x}}_i}^{\mathrm{T}} + \hat{\mathbf{F}}_{\boldsymbol{y}_i} \mathbf{\Sigma}_{\boldsymbol{y}_i} \hat{\mathbf{F}}_{\boldsymbol{y}_i}^{\mathrm{T}}$$

We now work out each of these four terms separately, making extensive use of the properties of the Kronecker product and columnwise outer product listed in the previous section:

$$\hat{\mathbf{F}}_{\hat{\boldsymbol{x}}_i} \mathbf{\Sigma}_{\hat{\boldsymbol{x}}_i} \hat{\mathbf{F}}_{\hat{\boldsymbol{x}}_i}^{\mathrm{T}} = (\hat{\boldsymbol{\mu}}_i \otimes \mathbf{W}_i \hat{\mathbf{A}})(\hat{\boldsymbol{\mu}}_i^{\mathrm{T}} \otimes \mathbf{\Sigma}_{\hat{\boldsymbol{x}}_i} \hat{\mathbf{A}}^{\mathrm{T}} \mathbf{W}_i)$$

$$
\begin{aligned}
&\quad + (\hat{\boldsymbol{\mu}}_i \otimes \mathbf{W}_i \hat{\mathbf{A}}) \left((\boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_i} - \boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_i} \hat{\mathbf{A}}^{\mathrm{T}} \mathbf{W}_i \hat{\boldsymbol{\Xi}}_i) \bowtie \mathbf{W}_i \boldsymbol{r}_i \right) \\
&\quad + \left(\mathbf{W}_i \boldsymbol{r}_i \bowtie (\mathbf{I}_{n_{\hat{x}}} - \hat{\mathbf{A}}^{\mathrm{T}} \mathbf{W}_i \hat{\boldsymbol{\Xi}}_i) \right) (\hat{\boldsymbol{\mu}}_i^{\mathrm{T}} \otimes \boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_i} \hat{\mathbf{A}}^{\mathrm{T}} \mathbf{W}_i) \\
&\quad + \left(\mathbf{W}_i \boldsymbol{r}_i \bowtie (\mathbf{I}_{n_{\hat{x}}} - \hat{\mathbf{A}}^{\mathrm{T}} \mathbf{W}_i \hat{\boldsymbol{\Xi}}_i) \right) \left((\boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_i} - \boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_i} \hat{\mathbf{A}}^{\mathrm{T}} \mathbf{W}_i \hat{\boldsymbol{\Xi}}_i) \bowtie \mathbf{W}_i \boldsymbol{r}_i \right) \\
&= \quad \hat{\boldsymbol{\mu}}_i \hat{\boldsymbol{\mu}}_i^{\mathrm{T}} \otimes \mathbf{W}_i \hat{\mathbf{A}} \boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_i} \hat{\mathbf{A}}^{\mathrm{T}} \mathbf{W}_i \\
&\quad + \mathbf{W}_i \hat{\mathbf{A}} (\boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_i} - \boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_i} \hat{\mathbf{A}}^{\mathrm{T}} \mathbf{W}_i \hat{\boldsymbol{\Xi}}_i) \bowtie \mathbf{W}_i \boldsymbol{r}_i \hat{\boldsymbol{\mu}}_i^{\mathrm{T}} \\
&\quad + \mathbf{W}_i \boldsymbol{r}_i \hat{\boldsymbol{\mu}}_i^{\mathrm{T}} \bowtie \mathbf{W}_i \hat{\mathbf{A}} (\boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_i} - \boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_i} \hat{\mathbf{A}}^{\mathrm{T}} \mathbf{W}_i \hat{\boldsymbol{\Xi}}_i) \\
&\quad + (\mathbf{I}_{n_{\hat{x}}} - \hat{\boldsymbol{\Xi}}_i^{\mathrm{T}} \mathbf{W}_i \hat{\mathbf{A}})(\boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_i} - \boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_i} \hat{\mathbf{A}}^{\mathrm{T}} \mathbf{W}_i \hat{\boldsymbol{\Xi}}_i) \otimes \mathbf{W}_i \boldsymbol{r}_i \boldsymbol{r}_i^{\mathrm{T}} \mathbf{W}_i
\end{aligned}
$$

$$
\begin{aligned}
\hat{\mathbf{F}}_{\hat{\boldsymbol{x}}_i} \boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_i \boldsymbol{y}_i}^{\mathrm{T}} \hat{\mathbf{F}}_{\boldsymbol{y}_i}^{\mathrm{T}} &= \quad - (\hat{\boldsymbol{\mu}}_i \otimes \mathbf{W}_i \hat{\mathbf{A}})(\hat{\boldsymbol{\mu}}_i^{\mathrm{T}} \otimes \boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_i \boldsymbol{y}_i}^{\mathrm{T}} \mathbf{W}_i) \\
&\quad + (\hat{\boldsymbol{\mu}}_i \otimes \mathbf{W}_i \hat{\mathbf{A}})(\boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_i \boldsymbol{y}_i}^{\mathrm{T}} \mathbf{W}_i \hat{\boldsymbol{\Xi}}_i \bowtie \mathbf{W}_i \boldsymbol{r}_i) \\
&\quad - \left(\mathbf{W}_i \boldsymbol{r}_i \bowtie (\mathbf{I}_{n_{\hat{x}}} - \hat{\mathbf{A}}^{\mathrm{T}} \mathbf{W}_i \hat{\boldsymbol{\Xi}}_i) \right) (\hat{\boldsymbol{\mu}}_i^{\mathrm{T}} \otimes \boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_i \boldsymbol{y}_i}^{\mathrm{T}} \mathbf{W}_i) \\
&\quad + \left(\mathbf{W}_i \boldsymbol{r}_i \bowtie (\mathbf{I}_{n_{\hat{x}}} - \hat{\mathbf{A}}^{\mathrm{T}} \mathbf{W}_i \hat{\boldsymbol{\Xi}}_i) \right) (\boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_i \boldsymbol{y}_i}^{\mathrm{T}} \mathbf{W}_i \hat{\boldsymbol{\Xi}}_i \bowtie \mathbf{W}_i \boldsymbol{r}_i) \\
&= \quad - \hat{\boldsymbol{\mu}}_i \hat{\boldsymbol{\mu}}_i^{\mathrm{T}} \otimes \mathbf{W}_i \hat{\mathbf{A}} \, \boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_i \boldsymbol{y}_i}^{\mathrm{T}} \mathbf{W}_i \\
&\quad + \mathbf{W}_i \hat{\mathbf{A}} \, \boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_i \boldsymbol{y}_i}^{\mathrm{T}} \mathbf{W}_i \hat{\boldsymbol{\Xi}}_i \bowtie \mathbf{W}_i \boldsymbol{r}_i \hat{\boldsymbol{\mu}}_i^{\mathrm{T}} \\
&\quad - \mathbf{W}_i \boldsymbol{r}_i \hat{\boldsymbol{\mu}}_i^{\mathrm{T}} \bowtie (\mathbf{W}_i \boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_i \boldsymbol{y}_i} - \mathbf{W}_i \boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_i \boldsymbol{y}_i} \hat{\mathbf{A}}^{\mathrm{T}} \mathbf{W}_i \hat{\boldsymbol{\Xi}}_i) \\
&\quad + (\boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_i \boldsymbol{y}_i}^{\mathrm{T}} \mathbf{W}_i \hat{\boldsymbol{\Xi}}_i - \hat{\boldsymbol{\Xi}}_i^{\mathrm{T}} \mathbf{W}_i \hat{\mathbf{A}} \, \boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_i \boldsymbol{y}_i}^{\mathrm{T}} \mathbf{W}_i \hat{\boldsymbol{\Xi}}_i) \otimes \mathbf{W}_i \boldsymbol{r}_i \boldsymbol{r}_i^{\mathrm{T}} \mathbf{W}_i
\end{aligned}
$$

$$
\begin{aligned}
\hat{\mathbf{F}}_{\boldsymbol{y}_i} \boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_i \boldsymbol{y}_i} \hat{\mathbf{F}}_{\hat{\boldsymbol{x}}_i}^{\mathrm{T}} &= \quad \left(\hat{\mathbf{F}}_{\hat{\boldsymbol{x}}_i} \boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_i \boldsymbol{y}_i}^{\mathrm{T}} \hat{\mathbf{F}}_{\boldsymbol{y}_i}^{\mathrm{T}} \right)^{\mathrm{T}} \\
&= \quad - \hat{\boldsymbol{\mu}}_i \hat{\boldsymbol{\mu}}_i^{\mathrm{T}} \otimes \mathbf{W}_i \boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_i \boldsymbol{y}_i} \hat{\mathbf{A}}^{\mathrm{T}} \mathbf{W}_i \\
&\quad + \mathbf{W}_i \boldsymbol{r}_i \hat{\boldsymbol{\mu}}_i^{\mathrm{T}} \bowtie \mathbf{W}_i \hat{\mathbf{A}} \, \boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_i \boldsymbol{y}_i}^{\mathrm{T}} \mathbf{W}_i \hat{\boldsymbol{\Xi}}_i \\
&\quad - (\mathbf{W}_i \boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_i \boldsymbol{y}_i} - \mathbf{W}_i \boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_i \boldsymbol{y}_i} \hat{\mathbf{A}}^{\mathrm{T}} \mathbf{W}_i \hat{\boldsymbol{\Xi}}_i) \bowtie \mathbf{W}_i \boldsymbol{r}_i \hat{\boldsymbol{\mu}}_i^{\mathrm{T}} \\
&\quad + (\hat{\boldsymbol{\Xi}}_i^{\mathrm{T}} \mathbf{W}_i \boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_i \boldsymbol{y}_i} - \hat{\boldsymbol{\Xi}}_i^{\mathrm{T}} \mathbf{W}_i \boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_i \boldsymbol{y}_i} \hat{\mathbf{A}}^{\mathrm{T}} \mathbf{W}_i \hat{\boldsymbol{\Xi}}_i) \otimes \mathbf{W}_i \boldsymbol{r}_i \boldsymbol{r}_i^{\mathrm{T}} \mathbf{W}_i
\end{aligned}
$$

$$
\begin{aligned}
\hat{\mathbf{F}}_{\boldsymbol{y}_i} \boldsymbol{\Sigma}_{\boldsymbol{y}_i}^{\mathrm{T}} \hat{\mathbf{F}}_{\boldsymbol{y}_i}^{\mathrm{T}} &= \quad (\hat{\boldsymbol{\mu}}_i \otimes \mathbf{W}_i)(\hat{\boldsymbol{\mu}}_i^{\mathrm{T}} \otimes \boldsymbol{\Sigma}_{\boldsymbol{y}_i} \mathbf{W}_i) \\
&\quad - (\hat{\boldsymbol{\mu}}_i \otimes \mathbf{W}_i)(\boldsymbol{\Sigma}_{\boldsymbol{y}_i} \mathbf{W}_i \hat{\boldsymbol{\Xi}}_i \bowtie \mathbf{W}_i \boldsymbol{r}_i)) \\
&\quad - (\mathbf{W}_i \boldsymbol{r}_i \bowtie \mathbf{W}_i \hat{\boldsymbol{\Xi}}_i)(\hat{\boldsymbol{\mu}}_i^{\mathrm{T}} \otimes \boldsymbol{\Sigma}_{\boldsymbol{y}_i} \mathbf{W}_i) \\
&\quad + (\mathbf{W}_i \boldsymbol{r}_i \bowtie \mathbf{W}_i \hat{\boldsymbol{\Xi}}_i)(\boldsymbol{\Sigma}_{\boldsymbol{y}_i} \mathbf{W}_i \hat{\boldsymbol{\Xi}}_i \bowtie \mathbf{W}_i \boldsymbol{r}_i)) \\
&= \quad \hat{\boldsymbol{\mu}}_i \hat{\boldsymbol{\mu}}_i^{\mathrm{T}} \otimes \mathbf{W}_i \boldsymbol{\Sigma}_{\boldsymbol{y}_i} \mathbf{W}_i \\
&\quad - \mathbf{W}_i \boldsymbol{\Sigma}_{\boldsymbol{y}_i} \mathbf{W}_i \hat{\boldsymbol{\Xi}}_i \bowtie \mathbf{W}_i \boldsymbol{r}_i \hat{\boldsymbol{\mu}}_i^{\mathrm{T}} \\
&\quad - \mathbf{W}_i \boldsymbol{r}_i \hat{\boldsymbol{\mu}}_i^{\mathrm{T}} \bowtie \mathbf{W}_i \boldsymbol{\Sigma}_{\boldsymbol{y}_i} \mathbf{W}_i \hat{\boldsymbol{\Xi}}_i \\
&\quad + \hat{\boldsymbol{\Xi}}_i^{\mathrm{T}} \mathbf{W}_i \boldsymbol{\Sigma}_{\boldsymbol{y}_i} \mathbf{W}_i \hat{\boldsymbol{\Xi}}_i \otimes \mathbf{W}_i \boldsymbol{r}_i \boldsymbol{r}_i^{\mathrm{T}} \mathbf{W}_i .
\end{aligned}
$$

Collecting terms in the sum, we write schematically

$$
\begin{aligned}
&\hat{\mathbf{F}}_{\hat{\boldsymbol{x}}_i} \boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_i} \hat{\mathbf{F}}_{\hat{\boldsymbol{x}}_i}^{\mathrm{T}} + \hat{\mathbf{F}}_{\hat{\boldsymbol{x}}_i} \boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_i \boldsymbol{y}_i}^{\mathrm{T}} \hat{\mathbf{F}}_{\boldsymbol{y}_i}^{\mathrm{T}} + \hat{\mathbf{F}}_{\boldsymbol{y}_i} \boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_i \boldsymbol{y}_i} \hat{\mathbf{F}}_{\hat{\boldsymbol{x}}_i}^{\mathrm{T}} + \hat{\mathbf{F}}_{\boldsymbol{y}_i} \boldsymbol{\Sigma}_{\boldsymbol{y}_i} \hat{\mathbf{F}}_{\boldsymbol{y}_i}^{\mathrm{T}} \\
&= \quad \hat{\boldsymbol{\mu}}_i \hat{\boldsymbol{\mu}}_i^{\mathrm{T}} \otimes \mathbf{T}_1 + \mathbf{T}_2 \bowtie \mathbf{W}_i \boldsymbol{r}_i \hat{\boldsymbol{\mu}}_i^{\mathrm{T}} + \mathbf{W}_i \boldsymbol{r}_i \hat{\boldsymbol{\mu}}_i^{\mathrm{T}} \bowtie \mathbf{T}_3 + \mathbf{T}_4 \otimes \mathbf{W}_i \boldsymbol{r}_i \boldsymbol{r}_i^{\mathrm{T}} \mathbf{W}_i
\end{aligned}
$$

where

$$
\mathbf{T}_1 \quad = \quad \mathbf{W}_i \hat{\mathbf{A}} \boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_i} \hat{\mathbf{A}}^{\mathrm{T}} \mathbf{W}_i - \mathbf{W}_i \hat{\mathbf{A}} \, \boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_i \boldsymbol{y}_i}^{\mathrm{T}} \mathbf{W}_i - \mathbf{W}_i \boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_i \boldsymbol{y}_i} \hat{\mathbf{A}}^{\mathrm{T}} \mathbf{W}_i + \mathbf{W}_i \boldsymbol{\Sigma}_{\boldsymbol{y}_i} \mathbf{W}_i
$$

$$= \mathbf{W}_i \left[\hat{\mathbf{A}} \boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_i} \hat{\mathbf{A}}^{\mathrm{T}} - \hat{\mathbf{A}}\, \boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_i \boldsymbol{y}_i}^{\mathrm{T}} - \boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_i \boldsymbol{y}_i} \hat{\mathbf{A}}^{\mathrm{T}} + \boldsymbol{\Sigma}_{\boldsymbol{y}_i} \right] \mathbf{W}_i \quad = \quad \mathbf{W}_i$$

$$\mathbf{T}_2 = \mathbf{W}_i \hat{\mathbf{A}} (\boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_i} - \boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_i} \hat{\mathbf{A}}^{\mathrm{T}} \mathbf{W}_i \hat{\boldsymbol{\Xi}}_i) + \mathbf{W}_i \hat{\mathbf{A}}\, \boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_i \boldsymbol{y}_i}^{\mathrm{T}} \mathbf{W}_i \hat{\boldsymbol{\Xi}}_i$$
$$- (\mathbf{W}_i \boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_i \boldsymbol{y}_i} - \mathbf{W}_i \boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_i \boldsymbol{y}_i} \hat{\mathbf{A}}^{\mathrm{T}} \mathbf{W}_i \hat{\boldsymbol{\Xi}}_i) - \mathbf{W}_i \boldsymbol{\Sigma}_{\boldsymbol{y}_i} \mathbf{W}_i \hat{\boldsymbol{\Xi}}_i$$
$$= \mathbf{W}_i \hat{\mathbf{A}}\, \boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_i} - \mathbf{W}_i \hat{\mathbf{A}} \boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_i} \hat{\mathbf{A}}^{\mathrm{T}} \mathbf{W}_i \hat{\boldsymbol{\Xi}}_i + \mathbf{W}_i \hat{\mathbf{A}}\, \boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_i \boldsymbol{y}_i}^{\mathrm{T}} \mathbf{W}_i \hat{\boldsymbol{\Xi}}_i$$
$$- \mathbf{W}_i \boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_i \boldsymbol{y}_i} + \mathbf{W}_i \boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_i \boldsymbol{y}_i} \hat{\mathbf{A}}^{\mathrm{T}} \mathbf{W}_i \hat{\boldsymbol{\Xi}}_i - \mathbf{W}_i \boldsymbol{\Sigma}_{\boldsymbol{y}_i} \mathbf{W}_i \hat{\boldsymbol{\Xi}}_i$$
$$= \mathbf{W}_i (\hat{\mathbf{A}}\, \boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_i} - \boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_i \boldsymbol{y}_i}) - \mathbf{W}_i \left(\hat{\mathbf{A}} \boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_i} \hat{\mathbf{A}}^{\mathrm{T}} - \hat{\mathbf{A}}\, \boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_i \boldsymbol{y}_i}^{\mathrm{T}} - \boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_i \boldsymbol{y}_i} \hat{\mathbf{A}}^{\mathrm{T}} + \boldsymbol{\Sigma}_{\boldsymbol{y}_i} \right) \mathbf{W}_i \hat{\boldsymbol{\Xi}}_i$$
$$= \mathbf{W}_i (\hat{\mathbf{A}} \boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_i} - \boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_i \boldsymbol{y}_i}) - \mathbf{W}_i (\hat{\mathbf{A}} \boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_i} - \boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_i \boldsymbol{y}_i}) \quad = \quad \mathbf{0}$$

$$\mathbf{T}_3 = \mathbf{T}_2 = \mathbf{0}$$

$$\mathbf{T}_4 = (\mathbf{I}_{n_{\hat{x}}} - \hat{\boldsymbol{\Xi}}_i^{\mathrm{T}} \mathbf{W}_i \hat{\mathbf{A}})(\boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_i} - \boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_i} \hat{\mathbf{A}}^{\mathrm{T}} \mathbf{W}_i \hat{\boldsymbol{\Xi}}_i)$$
$$+ (\boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_i \boldsymbol{y}_i}^{\mathrm{T}} \mathbf{W}_i \hat{\boldsymbol{\Xi}}_i - \hat{\boldsymbol{\Xi}}_i^{\mathrm{T}} \mathbf{W}_i \hat{\mathbf{A}}\, \boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_i \boldsymbol{y}_i}^{\mathrm{T}} \mathbf{W}_i \hat{\boldsymbol{\Xi}}_i)$$
$$+ (\hat{\boldsymbol{\Xi}}_i^{\mathrm{T}} \mathbf{W}_i \boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_i \boldsymbol{y}_i} - \hat{\boldsymbol{\Xi}}_i^{\mathrm{T}} \mathbf{W}_i \boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_i \boldsymbol{y}_i} \hat{\mathbf{A}}^{\mathrm{T}} \mathbf{W}_i \hat{\boldsymbol{\Xi}}_i) + \hat{\boldsymbol{\Xi}}_i^{\mathrm{T}} \mathbf{W}_i \boldsymbol{\Sigma}_{\boldsymbol{y}_i} \mathbf{W}_i \hat{\boldsymbol{\Xi}}_i$$
$$= \boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_i} - \boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_i} \hat{\mathbf{A}}^{\mathrm{T}} \mathbf{W}_i \hat{\boldsymbol{\Xi}}_i - \hat{\boldsymbol{\Xi}}_i^{\mathrm{T}} \mathbf{W}_i \hat{\mathbf{A}} \boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_i} + \hat{\boldsymbol{\Xi}}_i^{\mathrm{T}} \mathbf{W}_i \hat{\mathbf{A}} \boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_i} \hat{\mathbf{A}}^{\mathrm{T}} \mathbf{W}_i \hat{\boldsymbol{\Xi}}_i$$
$$+ \boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_i \boldsymbol{y}_i}^{\mathrm{T}} \mathbf{W}_i \hat{\boldsymbol{\Xi}}_i - \hat{\boldsymbol{\Xi}}_i^{\mathrm{T}} \mathbf{W}_i \hat{\mathbf{A}}\, \boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_i \boldsymbol{y}_i}^{\mathrm{T}} \mathbf{W}_i \hat{\boldsymbol{\Xi}}_i + \hat{\boldsymbol{\Xi}}_i^{\mathrm{T}} \mathbf{W}_i \boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_i \boldsymbol{y}_i}$$
$$- \hat{\boldsymbol{\Xi}}_i^{\mathrm{T}} \mathbf{W}_i \boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_i \boldsymbol{y}_i} \hat{\mathbf{A}}^{\mathrm{T}} \mathbf{W}_i \hat{\boldsymbol{\Xi}}_i + \hat{\boldsymbol{\Xi}}_i^{\mathrm{T}} \mathbf{W}_i \boldsymbol{\Sigma}_{\boldsymbol{y}_i} \mathbf{W}_i \hat{\boldsymbol{\Xi}}_i$$
$$= \boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_i} - (\hat{\mathbf{A}} \boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_i} - \boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_i \boldsymbol{y}_i})^{\mathrm{T}} \mathbf{W}_i \hat{\boldsymbol{\Xi}}_i - \hat{\boldsymbol{\Xi}}_i^{\mathrm{T}} \mathbf{W}_i (\hat{\mathbf{A}} \boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_i} - \boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_i \boldsymbol{y}_i})$$
$$+ \hat{\boldsymbol{\Xi}}_i^{\mathrm{T}} \mathbf{W}_i \left[\hat{\mathbf{A}} \boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_i} \hat{\mathbf{A}}^{\mathrm{T}} - \hat{\mathbf{A}}\, \boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_i \boldsymbol{y}_i}^{\mathrm{T}} - \boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_i \boldsymbol{y}_i} \hat{\mathbf{A}}^{\mathrm{T}} + \boldsymbol{\Sigma}_{\boldsymbol{y}_i} \right] \mathbf{W}_i \hat{\boldsymbol{\Xi}}_i$$
$$= \boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_i} - \hat{\boldsymbol{\Xi}}_i^{\mathrm{T}} \mathbf{W}_i \hat{\boldsymbol{\Xi}}_i$$

Putting it all together, we arrive at

$$\hat{\mathbf{Q}} = \sum_i \left\{ \hat{\boldsymbol{\mu}}_i \hat{\boldsymbol{\mu}}_i^{\mathrm{T}} \otimes \mathbf{W}_i + \boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_i} \otimes \mathbf{W}_i \boldsymbol{r}_i \boldsymbol{r}_i^{\mathrm{T}} \mathbf{W}_i - \hat{\boldsymbol{\Xi}}_i^{\,\mathrm{T}} \mathbf{W}_i \hat{\boldsymbol{\Xi}}_i \otimes \mathbf{W}_i \boldsymbol{r}_i \boldsymbol{r}_i^{\mathrm{T}} \mathbf{W}_i \right\}. \tag{72}$$

8.3 Derivation of reduced-size Jacobian matrix J

We begin by writing (45) as

$$\mathbf{F}(\mathbf{A}) \quad \dot{=} \quad \sum_i \mathbf{W}_i\, \boldsymbol{r}_i \boldsymbol{\mu}_i^{\mathrm{T}} = \mathbf{0} \tag{73}$$

Taking the derivative of (73),

$$\mathbf{J} = \frac{\partial \mathbf{F}}{\partial \mathbf{A}} = \sum_i \left\{ (\boldsymbol{\mu}_i \boldsymbol{r}_i^{\mathrm{T}} \otimes \mathbf{I}_{n_y}) \frac{\partial \mathbf{W}_i}{\partial \mathbf{A}} + (\boldsymbol{\mu}_i \otimes \mathbf{W}_i) \frac{\partial \boldsymbol{r}_i}{\partial \mathbf{A}} + (\mathbf{W}_i \boldsymbol{r}_i \bowtie \mathbf{I}_{n_x}) \frac{\partial \boldsymbol{\mu}_i}{\partial \mathbf{A}} \right\}$$
$$\dot{=} \quad \mathbf{J}_1 + \mathbf{J}_2 + \mathbf{J}_3 \tag{74}$$

where for future convenience we assign names to each of the three terms.

$$\frac{\partial \mathbf{W}_i}{\partial \mathbf{A}} = -(\mathbf{W}_i \otimes \mathbf{W}_i) \frac{\partial \mathbf{W}_i^{-1}}{\partial \mathbf{A}}$$
$$= -(\mathbf{W}_i \otimes \mathbf{W}_i) \frac{\partial}{\partial \mathbf{A}} \left[\mathbf{A}\, \boldsymbol{\Sigma}_{\boldsymbol{x}_i} \mathbf{A}^{\mathrm{T}} - \mathbf{A}\, \boldsymbol{\Sigma}_{\boldsymbol{x}_i \boldsymbol{y}_i}^{\mathrm{T}} - \boldsymbol{\Sigma}_{\boldsymbol{x}_i \boldsymbol{y}_i}\, \mathbf{A}^{\mathrm{T}} + \boldsymbol{\Sigma}_{\boldsymbol{y}_i} \right]$$
$$= -(\mathbf{W}_i \otimes \mathbf{W}_i) \left[(\mathbf{A} \boldsymbol{\Sigma}_{\boldsymbol{x}_i} \otimes \mathbf{I}_{n_y}) + (\mathbf{A} \boldsymbol{\Sigma}_{\boldsymbol{x}_i} \bowtie \mathbf{I}_{n_y}) - (\boldsymbol{\Sigma}_{\boldsymbol{x}_i \boldsymbol{y}_i} \otimes \mathbf{I}_{n_y}) - (\boldsymbol{\Sigma}_{\boldsymbol{x}_i \boldsymbol{y}_i} \bowtie \mathbf{I}_{n_y}) \right]$$
$$= - \left[\mathbf{W}_i (\mathbf{A} \boldsymbol{\Sigma}_{\boldsymbol{x}_i} - \boldsymbol{\Sigma}_{\boldsymbol{x}_i \boldsymbol{y}_i}) \otimes \mathbf{W}_i + \mathbf{W}_i (\mathbf{A} \boldsymbol{\Sigma}_{\boldsymbol{x}_i} - \boldsymbol{\Sigma}_{\boldsymbol{x}_i \boldsymbol{y}_i}) \bowtie \mathbf{W}_i \right]. \tag{75}$$

$$\frac{\partial \boldsymbol{r}_i}{\partial \mathbf{A}} = \frac{\partial}{\partial \mathbf{A}}\left[\mathbf{A}\boldsymbol{x}_i + \boldsymbol{a}_0 - \boldsymbol{y}_i\right]$$

$$= \boldsymbol{x}_i^{\mathrm{T}} \otimes \mathbf{I}_{n_y} + \frac{\partial \boldsymbol{a}_0}{\partial \mathbf{A}}. \tag{76}$$

$$\frac{\partial \boldsymbol{a}_0}{\partial \mathbf{A}} = -\frac{\partial}{\partial \mathbf{A}}\left[\boldsymbol{\Omega}^{-1}\sum_j \mathbf{W}_j(\mathbf{A}\boldsymbol{x}_j - \boldsymbol{y}_j)\right]$$

$$= -\left[\left(\sum_j (\mathbf{A}\boldsymbol{x}_j - \boldsymbol{y}_j)^{\mathrm{T}}\mathbf{W}_j \otimes \mathbf{I}_{n_y}\right)\frac{\partial \boldsymbol{\Omega}^{-1}}{\partial \mathbf{A}}\right.$$

$$\left. + \sum_j \left((\mathbf{A}\boldsymbol{x}_j - \boldsymbol{y}_j)^{\mathrm{T}} \otimes \boldsymbol{\Omega}^{-1}\right)\frac{\partial \mathbf{W}_j}{\partial \mathbf{A}} + \sum_j \boldsymbol{x}_j^{\mathrm{T}} \otimes \boldsymbol{\Omega}^{-1}\mathbf{W}_j\right]. \tag{77}$$

$$\frac{\partial \boldsymbol{\Omega}^{-1}}{\partial \mathbf{A}} = -\left(\boldsymbol{\Omega}^{-1} \otimes \boldsymbol{\Omega}^{-1}\right)\frac{\partial \boldsymbol{\Omega}}{\partial \mathbf{A}}$$

$$= -\left(\boldsymbol{\Omega}^{-1} \otimes \boldsymbol{\Omega}^{-1}\right)\sum_i \frac{\partial \mathbf{W}_i}{\partial \mathbf{A}}$$

$$= \left(\boldsymbol{\Omega}^{-1} \otimes \boldsymbol{\Omega}^{-1}\right)\sum_i \left[\mathbf{W}_i(\mathbf{A}\boldsymbol{\Sigma}_{\boldsymbol{x}_i} - \boldsymbol{\Sigma}_{\boldsymbol{x}_i\boldsymbol{y}_i}) \otimes \mathbf{W}_i + \mathbf{W}_i(\mathbf{A}\boldsymbol{\Sigma}_{\boldsymbol{x}_i} - \boldsymbol{\Sigma}_{\boldsymbol{x}_i\boldsymbol{y}_i}) \bowtie \mathbf{W}_i\right]$$

$$= \sum_i \left[\boldsymbol{\Omega}^{-1}\mathbf{W}_i(\mathbf{A}\boldsymbol{\Sigma}_{\boldsymbol{x}_i} - \boldsymbol{\Sigma}_{\boldsymbol{x}_i\boldsymbol{y}_i}) \otimes \boldsymbol{\Omega}^{-1}\mathbf{W}_i \right.$$

$$\left. + \boldsymbol{\Omega}^{-1}\mathbf{W}_i(\mathbf{A}\boldsymbol{\Sigma}_{\boldsymbol{x}_i} - \boldsymbol{\Sigma}_{\boldsymbol{x}_i\boldsymbol{y}_i}) \bowtie \mathbf{W}_i\boldsymbol{\Omega}^{-1}\right]. \tag{78}$$

$$\frac{\partial \boldsymbol{\mu}_i}{\partial \mathbf{A}} = \frac{\partial}{\partial \mathbf{A}}\left[\boldsymbol{x}_i - (\mathbf{A}\boldsymbol{\Sigma}_{\boldsymbol{x}_i} - \boldsymbol{\Sigma}_{\boldsymbol{x}_i\boldsymbol{y}_i})^{\mathrm{T}}\mathbf{W}_i\boldsymbol{r}_i\right]$$

$$= -\left[\boldsymbol{\Sigma}_{\boldsymbol{x}_i} \bowtie \mathbf{W}_i\boldsymbol{r}_i + \left(\boldsymbol{r}_i^{\mathrm{T}} \otimes (\mathbf{A}\boldsymbol{\Sigma}_{\boldsymbol{x}_i} - \boldsymbol{\Sigma}_{\boldsymbol{x}_i\boldsymbol{y}_i})^{\mathrm{T}}\right)\frac{\partial \mathbf{W}_i}{\partial \mathbf{A}}\right.$$

$$\left. + (\mathbf{A}\boldsymbol{\Sigma}_{\boldsymbol{x}_i} - \boldsymbol{\Sigma}_{\boldsymbol{x}_i\boldsymbol{y}_i})^{\mathrm{T}}\mathbf{W}_i\frac{\partial \boldsymbol{r}_i}{\partial \mathbf{A}}\right]. \tag{79}$$

From here on, to reduce clutter, we will set

$$\boldsymbol{\Xi}_i \doteq (\mathbf{A}\boldsymbol{\Sigma}_{\boldsymbol{x}_i} - \boldsymbol{\Sigma}_{\boldsymbol{x}_i\boldsymbol{y}_i})$$
$$\boldsymbol{z}_i \doteq \mathbf{A}\boldsymbol{x}_i - \boldsymbol{y}_i = \boldsymbol{r}_i - \boldsymbol{a}_0.$$

In what follows we will use $\boldsymbol{\mu}_i = \boldsymbol{x}_i - \boldsymbol{\Xi}_i^{\mathrm{T}}\mathbf{W}_i\boldsymbol{r}_i$ and $\boldsymbol{z}_i + \boldsymbol{a}_0 = \boldsymbol{r}_i$ to help simplify the expressions.

The first term inside brackets in (77) is

$$\left(\sum_j \boldsymbol{z}_j^{\mathrm{T}}\mathbf{W}_j \otimes \mathbf{I}_{n_y}\right)\sum_i \left[\boldsymbol{\Omega}^{-1}\mathbf{W}_i\boldsymbol{\Xi}_i \otimes \boldsymbol{\Omega}^{-1}\mathbf{W}_i + \boldsymbol{\Omega}^{-1}\mathbf{W}_i\boldsymbol{\Xi}_i \bowtie \mathbf{W}_i\boldsymbol{\Omega}^{-1}\right]$$

$$= \sum_j \sum_i \left\{\boldsymbol{z}_j^{\mathrm{T}}\mathbf{W}_j\boldsymbol{\Omega}^{-1}\mathbf{W}_i\boldsymbol{\Xi}_i \otimes \boldsymbol{\Omega}^{-1}\mathbf{W}_i + \boldsymbol{\Omega}^{-1}\mathbf{W}_i\boldsymbol{\Xi}_i \bowtie \mathbf{W}_i\boldsymbol{\Omega}^{-1}\mathbf{W}_j\boldsymbol{z}_j\right\}$$

$$= \boldsymbol{\Omega}^{-1}\sum_j \sum_i \left\{\boldsymbol{z}_j^{\mathrm{T}}\mathbf{W}_j\boldsymbol{\Omega}^{-1}\mathbf{W}_i\boldsymbol{\Xi}_i \otimes \mathbf{W}_i + \mathbf{W}_i\boldsymbol{\Xi}_i \bowtie \mathbf{W}_i\boldsymbol{\Omega}^{-1}\mathbf{W}_j\boldsymbol{z}_j\right\}$$

$$= \boldsymbol{\Omega}^{-1}\sum_i \left\{\left(\sum_j \boldsymbol{z}_j^{\mathrm{T}}\mathbf{W}_j\boldsymbol{\Omega}^{-1}\right)\mathbf{W}_i\boldsymbol{\Xi}_i \otimes \mathbf{W}_i + \mathbf{W}_i\boldsymbol{\Xi}_i \bowtie \mathbf{W}_i\left(\boldsymbol{\Omega}^{-1}\sum_j \mathbf{W}_j\boldsymbol{z}_j\right)\right\}$$

$$= -\mathbf{\Omega}^{-1} \sum_i \left\{ \boldsymbol{a}_0^{\mathrm{T}} \mathbf{W}_i \mathbf{\Xi}_i \otimes \mathbf{W}_i + \mathbf{W}_i \mathbf{\Xi}_i \bowtie \mathbf{W}_i \boldsymbol{a}_0 \right\}.$$

The second term inside brackets in (77) is

$$-\sum_j \left(\boldsymbol{z}_j^{\mathrm{T}} \otimes \mathbf{\Omega}^{-1} \right) \left[\mathbf{W}_j \mathbf{\Xi}_i \otimes \mathbf{W}_j + \mathbf{W}_j \mathbf{\Xi}_i \bowtie \mathbf{W}_j \right]$$

$$= -\sum_j \left\{ \boldsymbol{z}_j^{\mathrm{T}} \mathbf{W}_j \mathbf{\Xi}_i \otimes \mathbf{\Omega}^{-1} \mathbf{W}_j + \mathbf{\Omega}^{-1} \mathbf{W}_j \mathbf{\Xi}_i \bowtie \mathbf{W}_j \boldsymbol{z}_j \right\}$$

$$= -\mathbf{\Omega}^{-1} \sum_j \left\{ \boldsymbol{z}_j^{\mathrm{T}} \mathbf{W}_j \mathbf{\Xi}_i \otimes \mathbf{W}_j + \mathbf{W}_j \mathbf{\Xi}_i \bowtie \mathbf{W}_j \boldsymbol{z}_j \right\}.$$

The third term inside brackets in (77) is

$$\sum_j \boldsymbol{x}_j^{\mathrm{T}} \otimes \mathbf{\Omega}^{-1} \mathbf{W}_j \quad = \quad \mathbf{\Omega}^{-1} \sum_j \boldsymbol{x}_j^{\mathrm{T}} \otimes \mathbf{W}_j.$$

Thus (76) expands into

$$
\begin{aligned}
\frac{\partial \boldsymbol{r}_i}{\partial \mathbf{A}} &= \boldsymbol{x}_i^{\mathrm{T}} \otimes \mathbf{I}_{n_y} + \mathbf{\Omega}^{-1} \sum_j \left\{ \boldsymbol{a}_0^{\mathrm{T}} \mathbf{W}_j \mathbf{\Xi}_i \otimes \mathbf{W}_j + \mathbf{W}_j \mathbf{\Xi}_i \bowtie \mathbf{W}_j \boldsymbol{a}_0 \right. \\
&\qquad \left. + \boldsymbol{z}_j^{\mathrm{T}} \mathbf{W}_j \mathbf{\Xi}_i \otimes \mathbf{W}_j + \mathbf{W}_j \mathbf{\Xi}_i \bowtie \mathbf{W}_j \boldsymbol{z}_j - \boldsymbol{x}_j^{\mathrm{T}} \otimes \mathbf{W}_j \right\} \\
&= \boldsymbol{x}_i^{\mathrm{T}} \otimes \mathbf{I}_{n_y} + \mathbf{\Omega}^{-1} \sum_j \left\{ \boldsymbol{r}_j^{\mathrm{T}} \mathbf{W}_j \mathbf{\Xi}_i \otimes \mathbf{W}_j - \boldsymbol{x}_j^{\mathrm{T}} \otimes \mathbf{W}_j \right. \\
&\qquad \left. + \mathbf{W}_j \mathbf{\Xi}_i \bowtie \mathbf{W}_j \boldsymbol{r}_j \right\} \\
&= \boldsymbol{x}_i^{\mathrm{T}} \otimes \mathbf{I}_{n_y} + \mathbf{\Omega}^{-1} \sum_j \left\{ \mathbf{W}_j \mathbf{\Xi}_i \bowtie \mathbf{W}_j \boldsymbol{r}_j - \boldsymbol{\mu}_j^{\mathrm{T}} \otimes \mathbf{W}_j \right\}. \quad (80)
\end{aligned}
$$

The first term inside brackets in (79) is

$$\mathbf{\Sigma}_{\boldsymbol{x}_i} \bowtie \mathbf{W}_i \boldsymbol{r}_i$$

which we leave alone. The second term inside brackets in (79) is

$$- \left(\boldsymbol{r}_i^{\mathrm{T}} \otimes \mathbf{\Xi}_i^{\mathrm{T}} \mathbf{W}_i \right) \left[\mathbf{W}_i \mathbf{\Xi}_i \otimes \mathbf{W}_i + \mathbf{W}_i \mathbf{\Xi}_i \bowtie \mathbf{W}_i \right]$$

$$= - \left[\boldsymbol{r}_i^{\mathrm{T}} \mathbf{W}_i \mathbf{\Xi}_i \otimes \mathbf{\Xi}_i^{\mathrm{T}} \mathbf{W}_i + \mathbf{\Xi}_i^{\mathrm{T}} \mathbf{W}_i \mathbf{\Xi}_i \bowtie \mathbf{W}_i \boldsymbol{r}_i \right].$$

The third term inside brackets in (79) is

$$\mathbf{\Xi}_i^{\mathrm{T}} \mathbf{W}_i \left\{ \boldsymbol{x}_i^{\mathrm{T}} \otimes \mathbf{I}_{n_y} + \mathbf{\Omega}^{-1} \sum_j \left[\mathbf{W}_j \mathbf{\Xi}_i \bowtie \mathbf{W}_j \boldsymbol{r}_j - \boldsymbol{\mu}_j^{\mathrm{T}} \otimes \mathbf{W}_j \right] \right\}$$

$$= \boldsymbol{x}_i^{\mathrm{T}} \otimes \mathbf{\Xi}_i^{\mathrm{T}} \mathbf{W}_i + \sum_j \left[\mathbf{\Xi}_i^{\mathrm{T}} \mathbf{W}_i \mathbf{\Omega}^{-1} \mathbf{W}_j \mathbf{\Xi}_i \bowtie \mathbf{W}_j \boldsymbol{r}_j - \boldsymbol{\mu}_j^{\mathrm{T}} \otimes \mathbf{\Xi}_i^{\mathrm{T}} \mathbf{W}_i \mathbf{\Omega}^{-1} \mathbf{W}_j \right].$$

Thus (79) expands into

$$
\begin{aligned}
\frac{\partial \boldsymbol{\mu}_i}{\partial \mathbf{A}} &= -\mathbf{\Sigma}_{\boldsymbol{x}_i} \bowtie \mathbf{W}_i \boldsymbol{r}_i + \boldsymbol{r}_i^{\mathrm{T}} \mathbf{W}_i \mathbf{\Xi}_i \otimes \mathbf{\Xi}_i^{\mathrm{T}} \mathbf{W}_i + \mathbf{\Xi}_i^{\mathrm{T}} \mathbf{W}_i \mathbf{\Xi}_i \bowtie \mathbf{W}_i \boldsymbol{r}_i \\
&\quad - \boldsymbol{x}_i^{\mathrm{T}} \otimes \mathbf{\Xi}_i^{\mathrm{T}} \mathbf{W}_i
\end{aligned}
$$

$$-\sum_j \left[\boldsymbol{\Xi}_i^{\mathrm{T}} \mathbf{W}_i \boldsymbol{\Omega}^{-1} \mathbf{W}_j \boldsymbol{\Xi}_i \bowtie \mathbf{W}_j \boldsymbol{r}_j - \boldsymbol{\mu}_j^{\mathrm{T}} \otimes \boldsymbol{\Xi}_i^{\mathrm{T}} \mathbf{W}_i \boldsymbol{\Omega}^{-1} \mathbf{W}_j \right]$$

$$
\begin{aligned}
= \ & -\boldsymbol{\Sigma}_{\boldsymbol{x}_i} \bowtie \mathbf{W}_i \boldsymbol{r}_i + \boldsymbol{\Xi}_i^{\mathrm{T}} \mathbf{W}_i \boldsymbol{\Xi}_i \bowtie \mathbf{W}_i \boldsymbol{r}_i - \boldsymbol{\mu}_i^{\mathrm{T}} \otimes \boldsymbol{\Xi}_i^{\mathrm{T}} \mathbf{W}_i \\
& -\sum_j \left[\boldsymbol{\Xi}_i^{\mathrm{T}} \mathbf{W}_i \boldsymbol{\Omega}^{-1} \mathbf{W}_j \boldsymbol{\Xi}_i \bowtie \mathbf{W}_j \boldsymbol{r}_j - \boldsymbol{\mu}_j^{\mathrm{T}} \otimes \boldsymbol{\Xi}_i^{\mathrm{T}} \mathbf{W}_i \boldsymbol{\Omega}^{-1} \mathbf{W}_j \right].
\end{aligned}
\tag{81}
$$

The terms in (74) are

$$
\begin{aligned}
\mathbf{J}_1 \ = \ & \sum_i (\boldsymbol{\mu}_i \boldsymbol{r}_i^{\mathrm{T}} \otimes \mathbf{I}_{n_y}) \frac{\partial \mathbf{W}_i}{\partial \mathbf{A}} \\
= \ & \sum_i -(\boldsymbol{\mu}_i \boldsymbol{r}_i^{\mathrm{T}} \otimes \mathbf{I}_{n_y}) \left[\mathbf{W}_i \boldsymbol{\Xi}_i \otimes \mathbf{W}_i + \mathbf{W}_i \boldsymbol{\Xi}_i \bowtie \mathbf{W}_i \right] \\
= \ & -\sum_i \left\{ \boldsymbol{\mu}_i \boldsymbol{r}_i^{\mathrm{T}} \mathbf{W}_i \boldsymbol{\Xi}_i \otimes \mathbf{W}_i + \mathbf{W}_i \boldsymbol{\Xi}_i \bowtie \mathbf{W}_i \boldsymbol{r}_i \boldsymbol{\mu}_i^{\mathrm{T}} \right\}.
\end{aligned}
\tag{82}
$$

$$
\begin{aligned}
\mathbf{J}_2 \ = \ & \sum_i (\boldsymbol{\mu}_i \otimes \mathbf{W}_i) \frac{\partial \boldsymbol{r}_i}{\partial \mathbf{A}} \\
= \ & \sum_i (\boldsymbol{\mu}_i \otimes \mathbf{W}_i) \left\{ \boldsymbol{x}_i^{\mathrm{T}} \otimes \mathbf{I}_{n_y} + \boldsymbol{\Omega}^{-1} \sum_j \left[\mathbf{W}_j \boldsymbol{\Xi}_i \bowtie \mathbf{W}_j \boldsymbol{r}_j - \boldsymbol{\mu}_j^{\mathrm{T}} \otimes \mathbf{W}_j \right] \right\} \\
= \ & \sum_i \Big\{ \boldsymbol{\mu}_i \boldsymbol{x}_i^{\mathrm{T}} \otimes \mathbf{W}_i \\
& + \sum_j \left[\mathbf{W}_i \boldsymbol{\Omega}^{-1} \mathbf{W}_j \boldsymbol{\Xi}_i \bowtie \mathbf{W}_j \boldsymbol{r}_j \boldsymbol{\mu}_i^{\mathrm{T}} - \boldsymbol{\mu}_i \boldsymbol{\mu}_j^{\mathrm{T}} \otimes \mathbf{W}_i \boldsymbol{\Omega}^{-1} \mathbf{W}_j \right] \Big\}.
\end{aligned}
\tag{83}
$$

$$
\begin{aligned}
\mathbf{J}_3 \ = \ & \sum_i (\mathbf{W}_i \boldsymbol{r}_i \bowtie \mathbf{I}_{n_x}) \frac{\partial \boldsymbol{\mu}_i}{\partial \mathbf{A}} \\
= \ & \sum_i (\mathbf{W}_i \boldsymbol{r}_i \bowtie \mathbf{I}_{n_x}) \Big\{ -\boldsymbol{\Sigma}_{\boldsymbol{x}_i} \bowtie \mathbf{W}_i \boldsymbol{r}_i + \boldsymbol{\Xi}_i^{\mathrm{T}} \mathbf{W}_i \boldsymbol{\Xi}_i \bowtie \mathbf{W}_i \boldsymbol{r}_i - \boldsymbol{\mu}_i^{\mathrm{T}} \otimes \boldsymbol{\Xi}_i^{\mathrm{T}} \mathbf{W}_i \\
& -\sum_j \left[\boldsymbol{\Xi}_i^{\mathrm{T}} \mathbf{W}_i \boldsymbol{\Omega}^{-1} \mathbf{W}_j \boldsymbol{\Xi}_i \bowtie \mathbf{W}_j \boldsymbol{r}_j - \boldsymbol{\mu}_j^{\mathrm{T}} \otimes \boldsymbol{\Xi}_i^{\mathrm{T}} \mathbf{W}_i \boldsymbol{\Omega}^{-1} \mathbf{W}_j \right] \Big\} \\
= \ & \sum_i \Big\{ -\boldsymbol{\Sigma}_{\boldsymbol{x}_i} \otimes \mathbf{W}_i \boldsymbol{r}_i \boldsymbol{r}_i^{\mathrm{T}} \mathbf{W}_i \\
& + \boldsymbol{\Xi}_i^{\mathrm{T}} \mathbf{W}_i \boldsymbol{\Xi}_i \otimes \mathbf{W}_i \boldsymbol{r}_i \boldsymbol{r}_i^{\mathrm{T}} \mathbf{W}_i - \mathbf{W}_i \boldsymbol{r}_i \boldsymbol{\mu}_i^{\mathrm{T}} \bowtie \mathbf{W}_i \boldsymbol{\Xi}_i \\
& + \sum_j \left[\mathbf{W}_i \boldsymbol{r}_i \boldsymbol{\mu}_j^{\mathrm{T}} \bowtie \mathbf{W}_j \boldsymbol{\Omega}^{-1} \mathbf{W}_i \boldsymbol{\Xi}_i \right. \\
& \left. - \boldsymbol{\Xi}_i^{\mathrm{T}} \mathbf{W}_i \boldsymbol{\Omega}^{-1} \mathbf{W}_j \boldsymbol{\Xi}_i \otimes \mathbf{W}_i \boldsymbol{r}_i \boldsymbol{r}_j^{\mathrm{T}} \mathbf{W}_j \right] \Big\}.
\end{aligned}
\tag{84}
$$

Putting it all together,

$$
\begin{aligned}
\mathbf{J} \ = \ & \mathbf{J}_1 + \mathbf{J}_2 + \mathbf{J}_3 \\
= \ & \sum_i \Big\{ -\boldsymbol{\mu}_i \boldsymbol{r}_i^{\mathrm{T}} \mathbf{W}_i \boldsymbol{\Xi}_i \otimes \mathbf{W}_i - \mathbf{W}_i \boldsymbol{\Xi}_i \bowtie \mathbf{W}_i \boldsymbol{r}_i \boldsymbol{\mu}_i^{\mathrm{T}} \\
& + \boldsymbol{\mu}_i \boldsymbol{x}_i^{\mathrm{T}} \otimes \mathbf{W}_i \\
& - \boldsymbol{\Sigma}_{\boldsymbol{x}_i} \otimes \mathbf{W}_i \boldsymbol{r}_i \boldsymbol{r}_i^{\mathrm{T}} \mathbf{W}_i \\
& + \boldsymbol{\Xi}_i^{\mathrm{T}} \mathbf{W}_i \boldsymbol{\Xi}_i \otimes \mathbf{W}_i \boldsymbol{r}_i \boldsymbol{r}_i^{\mathrm{T}} \mathbf{W}_i - \mathbf{W}_i \boldsymbol{r}_i \boldsymbol{\mu}_i^{\mathrm{T}} \bowtie \mathbf{W}_i \boldsymbol{\Xi}_i \\
& + \sum_j \left[\mathbf{W}_i \boldsymbol{\Omega}^{-1} \mathbf{W}_j \boldsymbol{\Xi}_i \bowtie \mathbf{W}_j \boldsymbol{r}_j \boldsymbol{\mu}_i^{\mathrm{T}} - \boldsymbol{\mu}_i \boldsymbol{\mu}_j^{\mathrm{T}} \otimes \mathbf{W}_i \boldsymbol{\Omega}^{-1} \mathbf{W}_j \right.
\end{aligned}
$$

$$
\begin{aligned}
&\quad + \mathbf{W}_i \boldsymbol{r}_i \boldsymbol{\mu}_j^{\mathrm{T}} \bowtie \mathbf{W}_j \boldsymbol{\Omega}^{-1} \mathbf{W}_i \boldsymbol{\Xi}_i \\
&\quad - \boldsymbol{\Xi}_i^{\mathrm{T}} \mathbf{W}_i \boldsymbol{\Omega}^{-1} \mathbf{W}_j \boldsymbol{\Xi}_i \otimes \mathbf{W}_i \boldsymbol{r}_i \boldsymbol{r}_j^{\mathrm{T}} \mathbf{W}_j \big] \big\} \\
= \quad & \sum_i \big\{ \boldsymbol{\mu}_i \boldsymbol{\mu}_i^{\mathrm{T}} \otimes \mathbf{W}_i - \boldsymbol{\Sigma}_{\boldsymbol{x}_i} \otimes \mathbf{W}_i \boldsymbol{r}_i \boldsymbol{r}_i^{\mathrm{T}} \mathbf{W}_i \\
&\quad + \boldsymbol{\Xi}_i^{\mathrm{T}} \mathbf{W}_i \boldsymbol{\Xi}_i \otimes \mathbf{W}_i \boldsymbol{r}_i \boldsymbol{r}_i^{\mathrm{T}} \mathbf{W}_i \\
&\quad - \mathbf{W}_i \boldsymbol{\Xi}_i \bowtie \mathbf{W}_i \boldsymbol{r}_i \boldsymbol{\mu}_i^{\mathrm{T}} - \mathbf{W}_i \boldsymbol{r}_i \boldsymbol{\mu}_i^{\mathrm{T}} \bowtie \mathbf{W}_i \boldsymbol{\Xi}_i \\
&\quad + \sum_j \big[\mathbf{W}_i \boldsymbol{\Omega}^{-1} \mathbf{W}_j \boldsymbol{\Xi}_i \bowtie \mathbf{W}_j \boldsymbol{r}_j \boldsymbol{\mu}_i^{\mathrm{T}} + \mathbf{W}_i \boldsymbol{r}_i \boldsymbol{\mu}_j^{\mathrm{T}} \bowtie \mathbf{W}_j \boldsymbol{\Omega}^{-1} \mathbf{W}_i \boldsymbol{\Xi}_i \\
&\quad - \boldsymbol{\mu}_i \boldsymbol{\mu}_j^{\mathrm{T}} \otimes \mathbf{W}_i \boldsymbol{\Omega}^{-1} \mathbf{W}_j - \boldsymbol{\Xi}_i^{\mathrm{T}} \mathbf{W}_i \boldsymbol{\Omega}^{-1} \mathbf{W}_j \boldsymbol{\Xi}_i \otimes \mathbf{W}_i \boldsymbol{r}_i \boldsymbol{r}_j^{\mathrm{T}} \mathbf{W}_j \big] \big\} . \qquad (85)
\end{aligned}
$$

Bearing in mind that the order of summations can be interchanged, it is evident from the form of this expression that this matrix is symmetric. The portion before the sum over j is identical to (31) except for removing the hats on all quantities that take them. Now, the double sum can be factored into independent sums:

$$
\begin{aligned}
& \left(\sum_i \boldsymbol{\mu}_i \otimes \mathbf{W}_i \right) \boldsymbol{\Omega}^{-1} \left(\sum_j \mathbf{W}_j \boldsymbol{\Xi}_i \bowtie \mathbf{W}_j \boldsymbol{r}_j \right) \\
& + \left(\sum_i \mathbf{W}_i \boldsymbol{r}_i \bowtie \mathbf{W}_i \boldsymbol{\Xi}_i \right) \boldsymbol{\Omega}^{-1} \left(\sum_j \boldsymbol{\mu}_j^{\mathrm{T}} \otimes \mathbf{W}_j \right) \\
& - \left(\sum_i \boldsymbol{\mu}_i \otimes \mathbf{W}_i \right) \boldsymbol{\Omega}^{-1} \left(\sum_j \boldsymbol{\mu}_j^{\mathrm{T}} \otimes \mathbf{W}_j \right) \\
& - \left(\sum_i \boldsymbol{\Xi}_i^{\mathrm{T}} \mathbf{W}_i \otimes \mathbf{W}_i \boldsymbol{r}_i \right) \boldsymbol{\Omega}^{-1} \left(\sum_j \mathbf{W}_j \boldsymbol{\Xi}_i \otimes \boldsymbol{r}_j^{\mathrm{T}} \mathbf{W}_j \right) \qquad (86)
\end{aligned}
$$

Define

$$
\begin{aligned}
\mathbf{M}_i & \doteq \boldsymbol{\mu}_i \otimes \mathbf{I}_{n_y} \quad (n_x n_y \times n_y) & (87) \\
\mathbf{E}_i & \doteq \mathbf{W}_i \boldsymbol{r}_i \bowtie \boldsymbol{\Xi}_i = \boldsymbol{\Xi}_i^{\mathrm{T}} \otimes \mathbf{W}_i \boldsymbol{r}_i \quad (n_x n_y \times n_y) & (88)
\end{aligned}
$$

Then define

$$
\begin{aligned}
\langle \mathbf{M} \rangle & \doteq \sum_i \mathbf{M}_i \mathbf{W}_i \boldsymbol{\Omega}^{-1} = \left(\sum_i \boldsymbol{\mu}_i \otimes \mathbf{W}_i \right) \boldsymbol{\Omega}^{-1} & (89) \\
\langle \mathbf{E} \rangle & \doteq \sum_i \mathbf{E}_i \mathbf{W}_i \boldsymbol{\Omega}^{-1} = \left(\sum_i \boldsymbol{\Xi}_i^{\mathrm{T}} \mathbf{W}_i \otimes \mathbf{W}_i \boldsymbol{r}_i \right) \boldsymbol{\Omega}^{-1} & (90)
\end{aligned}
$$

Then this portion of $\mathbf{J}$ is

$$
\begin{aligned}
& \langle \mathbf{M} \rangle \boldsymbol{\Omega} \langle \mathbf{E} \rangle^{\mathrm{T}} + \langle \mathbf{E} \rangle \boldsymbol{\Omega} \langle \mathbf{M} \rangle^{\mathrm{T}} - \langle \mathbf{M} \rangle \boldsymbol{\Omega} \langle \mathbf{M} \rangle^{\mathrm{T}} - \langle \mathbf{E} \rangle \boldsymbol{\Omega} \langle \mathbf{E} \rangle^{\mathrm{T}} \\
& = - \left(\langle \mathbf{M} \rangle - \langle \mathbf{E} \rangle \right) \boldsymbol{\Omega} \left(\langle \mathbf{M} \rangle - \langle \mathbf{E} \rangle \right)^{\mathrm{T}}
\end{aligned}
$$

There does not seem to be any way to reduce $(\mathbf{M}_i - \mathbf{E}_i)$ to a single Kronecker product.

We finally obtain

$$
\begin{aligned}
\mathbf{J} \;=\;& \sum_i \big\{ \boldsymbol{\mu}_i \boldsymbol{\mu}_i^{\mathrm{T}} \otimes \mathbf{W}_i - \boldsymbol{\Sigma}_{\boldsymbol{x}_i} \otimes \mathbf{W}_i \boldsymbol{r}_i \boldsymbol{r}_i^{\mathrm{T}} \mathbf{W}_i \\
& + \boldsymbol{\Xi}_i^{\mathrm{T}} \mathbf{W}_i \boldsymbol{\Xi}_i \otimes \mathbf{W}_i \boldsymbol{r}_i \boldsymbol{r}_i^{\mathrm{T}} \mathbf{W}_i \\
& - \mathbf{W}_i \boldsymbol{\Xi}_i \bowtie \mathbf{W}_i \boldsymbol{r}_i \boldsymbol{\mu}_i^{\mathrm{T}} - \mathbf{W}_i \boldsymbol{r}_i \boldsymbol{\mu}_i^{\mathrm{T}} \bowtie \mathbf{W}_i \boldsymbol{\Xi}_i \big\} \\
& - \left(\langle \mathbf{M} \rangle - \langle \mathbf{E} \rangle \right) \boldsymbol{\Omega} \left(\langle \mathbf{M} \rangle - \langle \mathbf{E} \rangle \right)^{\mathrm{T}}
\end{aligned}
\tag{91}
$$

8.4   Derivation of reduced-size matrix Q

The covariance matrix for the full coefficient matrix $\mathbf{A}$ unpacks as

$$
\boldsymbol{\Sigma}_{\hat{\mathbf{A}}} \;\doteq\;
\begin{array}{c}
 n_y \\
 n_x n_y
\end{array}
\overset{\overset{n_y \quad\quad n_x n_y}{}}{
\left[
\begin{array}{cc}
\boldsymbol{\Sigma}_{\boldsymbol{a}_0} & \boldsymbol{\Sigma}_{\boldsymbol{a}_0,\mathbf{A}}^{\mathrm{T}} \\
\boldsymbol{\Sigma}_{\boldsymbol{a}_0,\mathbf{A}} & \boldsymbol{\Sigma}_{\mathbf{A}}
\end{array}
\right]}
\tag{92}
$$

$$
= \sum_i
\left[
\begin{array}{c}
\frac{\mathrm{d}\boldsymbol{a}_0}{\mathrm{d}\mathbbm{x}_i} \\
\frac{\mathrm{d}\mathbf{A}}{\mathrm{d}\mathbbm{x}_i}
\end{array}
\right]
\Sigma_i
\left[
\left(\frac{\mathrm{d}\boldsymbol{a}_0}{\mathrm{d}\mathbbm{x}_i} \right)^{\mathrm{T}}
\quad
\left(\frac{\mathrm{d}\mathbf{A}}{\mathrm{d}\mathbbm{x}_i} \right)^{\mathrm{T}}
\right]
$$

$$
=
\left[
\begin{array}{cc}
\sum_i \frac{\mathrm{d}\boldsymbol{a}_0}{\mathrm{d}\mathbbm{x}_i} \Sigma_i \left(\frac{\mathrm{d}\boldsymbol{a}_0}{\mathrm{d}\mathbbm{x}_i} \right)^{\mathrm{T}} & \sum_i \frac{\mathrm{d}\boldsymbol{a}_0}{\mathrm{d}\mathbbm{x}_i} \Sigma_i \left(\frac{\mathrm{d}\mathbf{A}}{\mathrm{d}\mathbbm{x}_i} \right)^{\mathrm{T}} \\
\sum_i \frac{\mathrm{d}\mathbf{A}}{\mathrm{d}\mathbbm{x}_i} \Sigma_i \left(\frac{\mathrm{d}\boldsymbol{a}_0}{\mathrm{d}\mathbbm{x}_i} \right)^{\mathrm{T}} & \sum_i \frac{\mathrm{d}\mathbf{A}}{\mathrm{d}\mathbbm{x}_i} \Sigma_i \left(\frac{\mathrm{d}\mathbf{A}}{\mathrm{d}\mathbbm{x}_i} \right)^{\mathrm{T}}
\end{array}
\right]
\tag{93}
$$

Following the same procedure as in Sec. 3.2, the covariance matrix for $\mathbf{A}$ is given by

$$
\boldsymbol{\Sigma}_{\mathbf{A}} \;=\; \mathbf{J}^{-1}\, \mathbf{Q}\, \mathbf{J}^{-1}
\tag{94}
$$

where

$$
\mathbf{Q} \;\doteq\; \sum_i \big\{ \mathbf{F}_{\boldsymbol{x}_i} \boldsymbol{\Sigma}_{\boldsymbol{x}_i} \mathbf{F}_{\boldsymbol{x}_i}^{\mathrm{T}} + \mathbf{F}_{\boldsymbol{x}_i} \boldsymbol{\Sigma}_{\boldsymbol{x}_i \boldsymbol{y}_i}^{\mathrm{T}} \mathbf{F}_{\boldsymbol{y}_i}^{\mathrm{T}} + \mathbf{F}_{\boldsymbol{y}_i} \boldsymbol{\Sigma}_{\boldsymbol{x}_i \boldsymbol{y}_i} \mathbf{F}_{\boldsymbol{x}_i}^{\mathrm{T}} + \mathbf{F}_{\boldsymbol{y}_i} \boldsymbol{\Sigma}_{\boldsymbol{y}_i} \mathbf{F}_{\boldsymbol{y}_i}^{\mathrm{T}} \big\}.
\tag{95}
$$

Now,

$$
\frac{\partial \mathbf{W}_j}{\partial \boldsymbol{x}_i} \;=\; \frac{\partial \mathbf{W}_j}{\partial \boldsymbol{y}_i} \;=\; \mathbf{0}, \quad i,j = 1,\dots,m
$$

$$
\frac{\partial \boldsymbol{a}_0}{\partial \boldsymbol{x}_i} \;=\; -\frac{\partial}{\partial \boldsymbol{x}_i}\left(\boldsymbol{\Omega}^{-1} \sum_j \mathbf{W}_j (\mathbf{A}\boldsymbol{x}_j - \boldsymbol{y}_j) \right) \;=\; -\boldsymbol{\Omega}^{-1} \mathbf{W}_i \mathbf{A}, \quad i = 1,\dots,m
$$

$$
\frac{\partial \boldsymbol{a}_0}{\partial \boldsymbol{y}_i} \;=\; -\frac{\partial}{\partial \boldsymbol{y}_i}\left(\boldsymbol{\Omega}^{-1} \sum_j \mathbf{W}_j (\mathbf{A}\boldsymbol{x}_j - \boldsymbol{y}_j) \right) \;=\; \boldsymbol{\Omega}^{-1} \mathbf{W}_i, \quad i = 1,\dots,m
$$

$$
\frac{\partial \boldsymbol{r}_j}{\partial \boldsymbol{x}_i} \;=\; \mathbf{A}\,\delta_{ij} + \frac{\partial \boldsymbol{a}_0}{\partial \boldsymbol{x}_i} \;=\; \mathbf{A}\,\delta_{ij} - \boldsymbol{\Omega}^{-1} \mathbf{W}_i \mathbf{A}, \quad i,j = 1,\dots,m
$$

$$
\frac{\partial \boldsymbol{r}_j}{\partial \boldsymbol{y}_i} \;=\; -\mathbf{I}_{n_y}\,\delta_{ij} + \frac{\partial \boldsymbol{a}_0}{\partial \boldsymbol{y}_i} \;=\; -\mathbf{I}_{n_y}\,\delta_{ij} + \boldsymbol{\Omega}^{-1} \mathbf{W}_i, \quad i,j = 1,\dots,m
$$

$$
\begin{aligned}
\frac{\partial \boldsymbol{\mu}_j}{\partial \boldsymbol{x}_i} \;&=\; \mathbf{I}_{n_x}\,\delta_{ij} - \boldsymbol{\Xi}_j^{\mathrm{T}} \mathbf{W}_j \frac{\partial \boldsymbol{r}_j}{\partial \boldsymbol{x}_i} \\
&=\; \mathbf{I}_{n_x}\,\delta_{ij} - \boldsymbol{\Xi}_i^{\mathrm{T}} \mathbf{W}_i \mathbf{A}\,\delta_{ij} + \boldsymbol{\Xi}_j^{\mathrm{T}} \mathbf{W}_j \boldsymbol{\Omega}^{-1} \mathbf{W}_i \mathbf{A}, \quad i,j = 1,\dots,m
\end{aligned}
$$

$$
\frac{\partial \boldsymbol{\mu}_j}{\partial \boldsymbol{y}_i} \;=\; -\boldsymbol{\Xi}_j^{\mathrm{T}} \mathbf{W}_j \frac{\partial \boldsymbol{r}_j}{\partial \boldsymbol{y}_i} \;=\; \boldsymbol{\Xi}_i^{\mathrm{T}} \mathbf{W}_i\,\delta_{ij} - \boldsymbol{\Xi}_j^{\mathrm{T}} \mathbf{W}_j \boldsymbol{\Omega}^{-1} \mathbf{W}_i, \quad i,j = 1,\dots,m
$$

then

$$
\begin{aligned}
\mathbf{F}_{\boldsymbol{x}_i} \;\; &= \;\; \frac{\partial \mathbf{F}}{\partial \boldsymbol{x}_i} \;\; = \;\; \frac{\partial}{\partial \boldsymbol{x}_i}\left(\sum_j \mathbf{W}_j\, \boldsymbol{r}_j \boldsymbol{\mu}_j^{\mathrm{T}}\right) \\
&= \;\; \sum_j \left\{ (\boldsymbol{\mu}_j \otimes \mathbf{W}_j)\, \frac{\partial \boldsymbol{r}_j}{\partial \boldsymbol{x}_i} + (\mathbf{W}_j\, \boldsymbol{r}_j \bowtie \mathbf{I}_{n_x})\, \frac{\partial \boldsymbol{\mu}_j}{\partial \boldsymbol{x}_i} \right\} \\
&= \;\; \sum_j \Big\{ (\boldsymbol{\mu}_j \otimes \mathbf{W}_j)\left(\mathbf{A}\,\delta_{ij} - \boldsymbol{\Omega}^{-1}\mathbf{W}_i\mathbf{A}\right) \\
&\qquad\quad + (\mathbf{W}_j\, \boldsymbol{r}_j \bowtie \mathbf{I}_{n_x})\left(\mathbf{I}_{n_x}\,\delta_{ij} - \boldsymbol{\Xi}_i^{\mathrm{T}}\mathbf{W}_i\mathbf{A}\,\delta_{ij} + \boldsymbol{\Xi}_j^{\mathrm{T}}\mathbf{W}_j\boldsymbol{\Omega}^{-1}\mathbf{W}_i\mathbf{A}\right)\Big\} \\
&= \;\; \boldsymbol{\mu}_i \otimes \mathbf{W}_i\mathbf{A} + \mathbf{W}_i\, \boldsymbol{r}_i \bowtie \left(\mathbf{I}_{n_x} - \mathbf{A}^{\mathrm{T}}\mathbf{W}_i\boldsymbol{\Xi}_i\right) \\
&\qquad - \sum_j \left\{ (\boldsymbol{\mu}_j \otimes \mathbf{W}_j) - (\mathbf{W}_j\, \boldsymbol{r}_j \bowtie \mathbf{W}_j\boldsymbol{\Xi}_i) \right\}\boldsymbol{\Omega}^{-1}\mathbf{W}_i\mathbf{A} \\
&= \;\; \boldsymbol{\mu}_i \otimes \mathbf{W}_i\mathbf{A} + \mathbf{W}_i\, \boldsymbol{r}_i \bowtie \left(\mathbf{I}_{n_x} - \mathbf{A}^{\mathrm{T}}\mathbf{W}_i\boldsymbol{\Xi}_i\right) - \langle \mathbf{Z} \rangle\, \mathbf{W}_i\mathbf{A} \qquad (96)
\end{aligned}
$$

where for convenience we have defined

$$
\langle \mathbf{Z} \rangle \;\; \doteq \;\; \langle \mathbf{M} \rangle - \langle \mathbf{E} \rangle \qquad (97)
$$

$$
\begin{aligned}
\mathbf{F}_{\boldsymbol{y}_i} \;\; &= \;\; \frac{\partial \mathbf{F}}{\partial \boldsymbol{y}_i} \;\; = \;\; \frac{\partial}{\partial \boldsymbol{y}_i}\left(\sum_j \mathbf{W}_j\, \boldsymbol{r}_j \boldsymbol{\mu}_j^{\mathrm{T}}\right) \\
&= \;\; \sum_j \left\{ (\boldsymbol{\mu}_j \otimes \mathbf{W}_j)\, \frac{\partial \boldsymbol{r}_j}{\partial \boldsymbol{y}_i} + (\mathbf{W}_j\, \boldsymbol{r}_j \bowtie \mathbf{I}_{n_x})\, \frac{\partial \boldsymbol{\mu}_j}{\partial \boldsymbol{y}_i} \right\} \\
&= \;\; \sum_j \Big\{ (\boldsymbol{\mu}_j \otimes \mathbf{W}_j)\left(-\mathbf{I}_{n_y}\,\delta_{ij} + \boldsymbol{\Omega}^{-1}\mathbf{W}_i\right) \\
&\qquad\quad + (\mathbf{W}_j\, \boldsymbol{r}_j \bowtie \mathbf{I}_{n_x})\left(\boldsymbol{\Xi}_i^{\mathrm{T}}\mathbf{W}_i\,\delta_{ij} - \boldsymbol{\Xi}_j^{\mathrm{T}}\mathbf{W}_j\boldsymbol{\Omega}^{-1}\mathbf{W}_i\right)\Big\} \\
&= \;\; -\boldsymbol{\mu}_i \otimes \mathbf{W}_i + \mathbf{W}_i\, \boldsymbol{r}_i \bowtie \mathbf{W}_i\boldsymbol{\Xi}_i \\
&\qquad + \sum_j \left\{ (\boldsymbol{\mu}_j \otimes \mathbf{W}_j) - (\mathbf{W}_j\, \boldsymbol{r}_j \bowtie \mathbf{W}_j\boldsymbol{\Xi}_i) \right\}\boldsymbol{\Omega}^{-1}\mathbf{W}_i \\
&= \;\; -\boldsymbol{\mu}_i \otimes \mathbf{W}_i + \mathbf{W}_i\, \boldsymbol{r}_i \bowtie \mathbf{W}_i\boldsymbol{\Xi}_i + \langle \mathbf{Z} \rangle\, \mathbf{W}_i \qquad (98)
\end{aligned}
$$

To avoid dizziness when collecting terms below, line numbers in brackets are attached to the final expression.

$$
\begin{aligned}
\mathbf{F}_{\boldsymbol{x}_i}\boldsymbol{\Sigma}_{\boldsymbol{x}_i}\mathbf{F}_{\boldsymbol{x}_i}^{\mathrm{T}} \;\; &= \;\; (\boldsymbol{\mu}_i \otimes \mathbf{W}_i\mathbf{A})\left(\boldsymbol{\mu}_i^{\mathrm{T}} \otimes \boldsymbol{\Sigma}_{\boldsymbol{x}_i}\mathbf{A}^{\mathrm{T}}\mathbf{W}_i\right) \\
&\quad + (\boldsymbol{\mu}_i \otimes \mathbf{W}_i\mathbf{A})\left((\boldsymbol{\Sigma}_{\boldsymbol{x}_i} - \boldsymbol{\Sigma}_{\boldsymbol{x}_i}\mathbf{A}^{\mathrm{T}}\mathbf{W}_i\boldsymbol{\Xi}_i) \bowtie \mathbf{W}_i\, \boldsymbol{r}_i\right) \\
&\quad - (\boldsymbol{\mu}_i \otimes \mathbf{W}_i\mathbf{A})\,\boldsymbol{\Sigma}_{\boldsymbol{x}_i}\mathbf{A}^{\mathrm{T}}\mathbf{W}_i\langle \mathbf{Z} \rangle^{\mathrm{T}} \\
&\quad + \left(\mathbf{W}_i\, \boldsymbol{r}_i \bowtie (\mathbf{I}_{n_x} - \mathbf{A}^{\mathrm{T}}\mathbf{W}_i\boldsymbol{\Xi}_i)\right)\left(\boldsymbol{\mu}_i^{\mathrm{T}} \otimes \boldsymbol{\Sigma}_{\boldsymbol{x}_i}\mathbf{A}^{\mathrm{T}}\mathbf{W}_i\right) \\
&\quad + \left(\mathbf{W}_i\, \boldsymbol{r}_i \bowtie (\mathbf{I}_{n_x} - \mathbf{A}^{\mathrm{T}}\mathbf{W}_i\boldsymbol{\Xi}_i)\right)\left((\boldsymbol{\Sigma}_{\boldsymbol{x}_i} - \boldsymbol{\Sigma}_{\boldsymbol{x}_i}\mathbf{A}^{\mathrm{T}}\mathbf{W}_i\boldsymbol{\Xi}_i) \bowtie \mathbf{W}_i\, \boldsymbol{r}_i\right) \\
&\quad - \left(\mathbf{W}_i\, \boldsymbol{r}_i \bowtie (\mathbf{I}_{n_x} - \mathbf{A}^{\mathrm{T}}\mathbf{W}_i\boldsymbol{\Xi}_i)\right)\boldsymbol{\Sigma}_{\boldsymbol{x}_i}\mathbf{A}^{\mathrm{T}}\mathbf{W}_i\langle \mathbf{Z} \rangle^{\mathrm{T}} \\
&\quad - \langle \mathbf{Z} \rangle\, \mathbf{W}_i\mathbf{A}\left(\boldsymbol{\mu}_i^{\mathrm{T}} \otimes \boldsymbol{\Sigma}_{\boldsymbol{x}_i}\mathbf{A}^{\mathrm{T}}\mathbf{W}_i\right) \\
&\quad - \langle \mathbf{Z} \rangle\, \mathbf{W}_i\mathbf{A}\left((\boldsymbol{\Sigma}_{\boldsymbol{x}_i} - \boldsymbol{\Sigma}_{\boldsymbol{x}_i}\mathbf{A}^{\mathrm{T}}\mathbf{W}_i\boldsymbol{\Xi}_i) \bowtie \mathbf{W}_i\, \boldsymbol{r}_i\right)
\end{aligned}
$$

25

$$+ \langle\mathbf{Z}\rangle\,\mathbf{W}_i\mathbf{A}\boldsymbol{\Sigma}_{\boldsymbol{x}_i}\mathbf{A}^{\mathrm{T}}\mathbf{W}_i\langle\mathbf{Z}\rangle^{\mathrm{T}}$$

$$
\begin{aligned}
=\quad & [1]\ \ \boldsymbol{\mu}_i\boldsymbol{\mu}_i^{\mathrm{T}}\otimes\mathbf{W}_i\mathbf{A}\boldsymbol{\Sigma}_{\boldsymbol{x}_i}\mathbf{A}^{\mathrm{T}}\mathbf{W}_i\\
& [2]\ +\mathbf{W}_i(\mathbf{A}\boldsymbol{\Sigma}_{\boldsymbol{x}_i}-\mathbf{A}\boldsymbol{\Sigma}_{\boldsymbol{x}_i}\mathbf{A}^{\mathrm{T}}\mathbf{W}_i\boldsymbol{\Xi}_i)\bowtie\mathbf{W}_i\,\boldsymbol{r}_i\boldsymbol{\mu}_i^{\mathrm{T}}\\
& [3]\ -\boldsymbol{\mu}_i\otimes\mathbf{W}_i\mathbf{A}\boldsymbol{\Sigma}_{\boldsymbol{x}_i}\mathbf{A}^{\mathrm{T}}\mathbf{W}_i\langle\mathbf{Z}\rangle^{\mathrm{T}}\\
& [4]\ +\mathbf{W}_i\,\boldsymbol{r}_i\boldsymbol{\mu}_i^{\mathrm{T}}\bowtie\mathbf{W}_i(\mathbf{A}\boldsymbol{\Sigma}_{\boldsymbol{x}_i}-\mathbf{A}\boldsymbol{\Sigma}_{\boldsymbol{x}_i}\mathbf{A}^{\mathrm{T}}\mathbf{W}_i\boldsymbol{\Xi}_i)\\
& [5]\ +(\mathbf{I}_{n_x}-\boldsymbol{\Xi}_i^{\mathrm{T}}\mathbf{W}_i\mathbf{A})(\boldsymbol{\Sigma}_{\boldsymbol{x}_i}-\boldsymbol{\Sigma}_{\boldsymbol{x}_i}\mathbf{A}^{\mathrm{T}}\mathbf{W}_i\boldsymbol{\Xi}_i)\otimes\mathbf{W}_i\,\boldsymbol{r}_i\boldsymbol{r}_i^{\mathrm{T}}\mathbf{W}_i\\
& [6]\ -\mathbf{W}_i\,\boldsymbol{r}_i\bowtie\langle\mathbf{Z}\rangle\mathbf{W}_i(\mathbf{A}\boldsymbol{\Sigma}_{\boldsymbol{x}_i}-\mathbf{A}\boldsymbol{\Sigma}_{\boldsymbol{x}_i}\mathbf{A}^{\mathrm{T}}\mathbf{W}_i\boldsymbol{\Xi}_i)\\
& [7]\ -\boldsymbol{\mu}_i^{\mathrm{T}}\otimes\langle\mathbf{Z}\rangle\,\mathbf{W}_i\mathbf{A}\boldsymbol{\Sigma}_{\boldsymbol{x}_i}\mathbf{A}^{\mathrm{T}}\mathbf{W}_i\\
& [8]\ -\langle\mathbf{Z}\rangle\,\mathbf{W}_i(\mathbf{A}\boldsymbol{\Sigma}_{\boldsymbol{x}_i}-\mathbf{A}\boldsymbol{\Sigma}_{\boldsymbol{x}_i}\mathbf{A}^{\mathrm{T}}\mathbf{W}_i\boldsymbol{\Xi}_i)\bowtie\mathbf{W}_i\,\boldsymbol{r}_i\\
& [9]\ +\langle\mathbf{Z}\rangle\,\mathbf{W}_i\mathbf{A}\boldsymbol{\Sigma}_{\boldsymbol{x}_i}\mathbf{A}^{\mathrm{T}}\mathbf{W}_i\langle\mathbf{Z}\rangle^{\mathrm{T}}
\end{aligned}
$$

The other terms in $\mathbf{Q}$ can be written down by appropriate substitution of $\mathbf{A}$ by -1, $\boldsymbol{\Sigma}_{\boldsymbol{x}_i}$ by $\boldsymbol{\Sigma}_{\boldsymbol{x}_i\boldsymbol{y}_i}$, $\boldsymbol{\Sigma}_{\boldsymbol{x}_i\boldsymbol{y}_i}^{\mathrm{T}}$ or $\boldsymbol{\Sigma}_{\boldsymbol{y}_i}$, and omitting of $\mathbf{I}_{n_x}$.

$$
\begin{aligned}
\mathbf{F}_{\boldsymbol{x}_i}\boldsymbol{\Sigma}_{\boldsymbol{x}_i\boldsymbol{y}_i}^{\mathrm{T}}\mathbf{F}_{\boldsymbol{y}_i}^{\mathrm{T}} =\quad & [1]\ -\boldsymbol{\mu}_i\boldsymbol{\mu}_i^{\mathrm{T}}\otimes\mathbf{W}_i\mathbf{A}\boldsymbol{\Sigma}_{\boldsymbol{x}_i\boldsymbol{y}_i}^{\mathrm{T}}\mathbf{W}_i\\
& [2]\ +\mathbf{W}_i\mathbf{A}\boldsymbol{\Sigma}_{\boldsymbol{x}_i\boldsymbol{y}_i}^{\mathrm{T}}\mathbf{W}_i\boldsymbol{\Xi}_i\bowtie\mathbf{W}_i\,\boldsymbol{r}_i\boldsymbol{\mu}_i^{\mathrm{T}}\\
& [3]\ +\boldsymbol{\mu}_i\otimes\mathbf{W}_i\mathbf{A}\boldsymbol{\Sigma}_{\boldsymbol{x}_i\boldsymbol{y}_i}^{\mathrm{T}}\mathbf{W}_i\langle\mathbf{Z}\rangle^{\mathrm{T}}\\
& [4]\ -\mathbf{W}_i\,\boldsymbol{r}_i\boldsymbol{\mu}_i^{\mathrm{T}}\bowtie\mathbf{W}_i(\boldsymbol{\Sigma}_{\boldsymbol{x}_i\boldsymbol{y}_i}-\boldsymbol{\Sigma}_{\boldsymbol{x}_i\boldsymbol{y}_i}\mathbf{A}^{\mathrm{T}}\mathbf{W}_i\boldsymbol{\Xi}_i)\\
& [5]\ +(\boldsymbol{\Sigma}_{\boldsymbol{x}_i\boldsymbol{y}_i}^{\mathrm{T}}\mathbf{W}_i\boldsymbol{\Xi}_i-\boldsymbol{\Xi}_i^{\mathrm{T}}\mathbf{W}_i\mathbf{A}\boldsymbol{\Sigma}_{\boldsymbol{x}_i\boldsymbol{y}_i}^{\mathrm{T}}\mathbf{W}_i\boldsymbol{\Xi}_i)\otimes\mathbf{W}_i\,\boldsymbol{r}_i\boldsymbol{r}_i^{\mathrm{T}}\mathbf{W}_i\\
& [6]\ +\mathbf{W}_i\,\boldsymbol{r}_i\bowtie\langle\mathbf{Z}\rangle\mathbf{W}_i(\boldsymbol{\Sigma}_{\boldsymbol{x}_i\boldsymbol{y}_i}-\boldsymbol{\Sigma}_{\boldsymbol{x}_i\boldsymbol{y}_i}\mathbf{A}^{\mathrm{T}}\mathbf{W}_i\boldsymbol{\Xi}_i)\\
& [7]\ +\boldsymbol{\mu}_i^{\mathrm{T}}\otimes\langle\mathbf{Z}\rangle\,\mathbf{W}_i\mathbf{A}\boldsymbol{\Sigma}_{\boldsymbol{x}_i\boldsymbol{y}_i}^{\mathrm{T}}\mathbf{W}_i\\
& [8]\ -\langle\mathbf{Z}\rangle\,\mathbf{W}_i\mathbf{A}\boldsymbol{\Sigma}_{\boldsymbol{x}_i\boldsymbol{y}_i}^{\mathrm{T}}\mathbf{W}_i\boldsymbol{\Xi}_i\bowtie\mathbf{W}_i\,\boldsymbol{r}_i\\
& [9]\ -\langle\mathbf{Z}\rangle\,\mathbf{W}_i\mathbf{A}\boldsymbol{\Sigma}_{\boldsymbol{x}_i\boldsymbol{y}_i}^{\mathrm{T}}\mathbf{W}_i\langle\mathbf{Z}\rangle^{\mathrm{T}}\\[4pt]
\mathbf{F}_{\boldsymbol{y}_i}\boldsymbol{\Sigma}_{\boldsymbol{x}_i\boldsymbol{y}_i}\mathbf{F}_{\boldsymbol{x}_i}^{\mathrm{T}} =\quad & (\mathbf{F}_{\boldsymbol{x}_i}\boldsymbol{\Sigma}_{\boldsymbol{x}_i\boldsymbol{y}_i}^{\mathrm{T}}\mathbf{F}_{\boldsymbol{y}_i}^{\mathrm{T}})^{\mathrm{T}}\\
=\quad & [1]\ -\boldsymbol{\mu}_i\boldsymbol{\mu}_i^{\mathrm{T}}\otimes\mathbf{W}_i\boldsymbol{\Sigma}_{\boldsymbol{x}_i\boldsymbol{y}_i}\mathbf{A}^{\mathrm{T}}\mathbf{W}_i\\
& [4]\ +\mathbf{W}_i\,\boldsymbol{r}_i\boldsymbol{\mu}_i^{\mathrm{T}}\bowtie\mathbf{W}_i\mathbf{A}\boldsymbol{\Sigma}_{\boldsymbol{x}_i\boldsymbol{y}_i}^{\mathrm{T}}\mathbf{W}_i\boldsymbol{\Xi}_i\\
& [7]\ +\boldsymbol{\mu}_i^{\mathrm{T}}\otimes\langle\mathbf{Z}\rangle\,\mathbf{W}_i\boldsymbol{\Sigma}_{\boldsymbol{x}_i\boldsymbol{y}_i}\mathbf{A}^{\mathrm{T}}\mathbf{W}_i\\
& [2]\ -\mathbf{W}_i(\boldsymbol{\Sigma}_{\boldsymbol{x}_i\boldsymbol{y}_i}-\boldsymbol{\Sigma}_{\boldsymbol{x}_i\boldsymbol{y}_i}\mathbf{A}^{\mathrm{T}}\mathbf{W}_i\boldsymbol{\Xi}_i)\bowtie\mathbf{W}_i\,\boldsymbol{r}_i\boldsymbol{\mu}_i^{\mathrm{T}}\\
& [5]\ +(\boldsymbol{\Xi}_i^{\mathrm{T}}\mathbf{W}_i\boldsymbol{\Sigma}_{\boldsymbol{x}_i\boldsymbol{y}_i}-\boldsymbol{\Xi}_i^{\mathrm{T}}\mathbf{W}_i\boldsymbol{\Sigma}_{\boldsymbol{x}_i\boldsymbol{y}_i}\mathbf{A}^{\mathrm{T}}\mathbf{W}_i\boldsymbol{\Xi}_i)\otimes\mathbf{W}_i\,\boldsymbol{r}_i\boldsymbol{r}_i^{\mathrm{T}}\mathbf{W}_i\\
& [8]\ +\langle\mathbf{Z}\rangle\mathbf{W}_i(\boldsymbol{\Sigma}_{\boldsymbol{x}_i\boldsymbol{y}_i}-\boldsymbol{\Sigma}_{\boldsymbol{x}_i\boldsymbol{y}_i}\mathbf{A}^{\mathrm{T}}\mathbf{W}_i\boldsymbol{\Xi}_i)\bowtie\mathbf{W}_i\,\boldsymbol{r}_i\\
& [3]\ +\boldsymbol{\mu}_i\otimes\mathbf{W}_i\boldsymbol{\Sigma}_{\boldsymbol{x}_i\boldsymbol{y}_i}\mathbf{A}^{\mathrm{T}}\mathbf{W}_i\langle\mathbf{Z}\rangle^{\mathrm{T}}\\
& [6]\ -\mathbf{W}_i\,\boldsymbol{r}_i\bowtie\langle\mathbf{Z}\rangle\,\mathbf{W}_i\mathbf{A}\boldsymbol{\Sigma}_{\boldsymbol{x}_i\boldsymbol{y}_i}^{\mathrm{T}}\mathbf{W}_i\boldsymbol{\Xi}_i\\
& [9]\ -\langle\mathbf{Z}\rangle\,\mathbf{W}_i\boldsymbol{\Sigma}_{\boldsymbol{x}_i\boldsymbol{y}_i}\mathbf{A}^{\mathrm{T}}\mathbf{W}_i\langle\mathbf{Z}\rangle^{\mathrm{T}}\\[4pt]
\mathbf{F}_{\boldsymbol{y}_i}\boldsymbol{\Sigma}_{\boldsymbol{y}_i}\mathbf{F}_{\boldsymbol{y}_i}^{\mathrm{T}} =\quad & [1]\ \ \boldsymbol{\mu}_i\boldsymbol{\mu}_i^{\mathrm{T}}\otimes\mathbf{W}_i\boldsymbol{\Sigma}_{\boldsymbol{y}_i}\mathbf{W}_i\\
& [2]\ -\mathbf{W}_i\boldsymbol{\Sigma}_{\boldsymbol{y}_i}\mathbf{W}_i\boldsymbol{\Xi}_i\bowtie\mathbf{W}_i\,\boldsymbol{r}_i\boldsymbol{\mu}_i^{\mathrm{T}}\\
& [3]\ -\boldsymbol{\mu}_i\otimes\mathbf{W}_i\boldsymbol{\Sigma}_{\boldsymbol{y}_i}\mathbf{W}_i\langle\mathbf{Z}\rangle^{\mathrm{T}}\\
& [4]\ -\mathbf{W}_i\,\boldsymbol{r}_i\boldsymbol{\mu}_i^{\mathrm{T}}\bowtie\mathbf{W}_i\boldsymbol{\Sigma}_{\boldsymbol{y}_i}\mathbf{W}_i\boldsymbol{\Xi}_i\\
& [5]\ +\boldsymbol{\Xi}_i^{\mathrm{T}}\mathbf{W}_i\boldsymbol{\Sigma}_{\boldsymbol{y}_i}\mathbf{W}_i\boldsymbol{\Xi}_i\otimes\mathbf{W}_i\,\boldsymbol{r}_i\boldsymbol{r}_i^{\mathrm{T}}\mathbf{W}_i\\
& [6]\ +\mathbf{W}_i\,\boldsymbol{r}_i\bowtie\langle\mathbf{Z}\rangle\mathbf{W}_i\boldsymbol{\Sigma}_{\boldsymbol{y}_i}\mathbf{W}_i\boldsymbol{\Xi}_i\\
& [7]\ -\boldsymbol{\mu}_i^{\mathrm{T}}\otimes\langle\mathbf{Z}\rangle\,\mathbf{W}_i\boldsymbol{\Sigma}_{\boldsymbol{y}_i}\mathbf{W}_i
\end{aligned}
$$

$$[8] \quad + \langle \mathbf{Z} \rangle \, \mathbf{W}_i \boldsymbol{\Sigma}_{\boldsymbol{y}_i} \mathbf{W}_i \boldsymbol{\Xi}_i \bowtie \mathbf{W}_i \, \boldsymbol{r}_i$$
$$[9] \quad + \langle \mathbf{Z} \rangle \, \mathbf{W}_i \boldsymbol{\Sigma}_{\boldsymbol{y}_i} \mathbf{W}_i \langle \mathbf{Z} \rangle^{\mathrm{T}}$$

Upon collecting terms, lines 1, 2, 4, and 5 are the same as for the full-size case and will yield the same result as (72) with the removal of all hats. The additional terms here are:

$$
\begin{aligned}
[3] \quad &= \quad -\boldsymbol{\mu}_i \otimes \mathbf{W}_i \left[\mathbf{A}\boldsymbol{\Sigma}_{\boldsymbol{x}_i} \mathbf{A}^{\mathrm{T}} - \mathbf{A}\boldsymbol{\Sigma}_{\boldsymbol{x}_i \boldsymbol{y}_i}^{\mathrm{T}} - \boldsymbol{\Sigma}_{\boldsymbol{x}_i \boldsymbol{y}_i} \mathbf{A}^{\mathrm{T}} + \boldsymbol{\Sigma}_{\boldsymbol{y}_i} \right] \mathbf{W}_i \langle \mathbf{Z} \rangle^{\mathrm{T}} \\
&= \quad -\boldsymbol{\mu}_i \otimes \mathbf{W}_i \langle \mathbf{Z} \rangle^{\mathrm{T}}
\end{aligned}
$$

and transposing gives

$$[7] \quad = \quad -\boldsymbol{\mu}_i^{\mathrm{T}} \otimes \langle \mathbf{Z} \rangle \mathbf{W}_i$$

$$
\begin{aligned}
[6] \quad &= \quad -\mathbf{W}_i \, \boldsymbol{r}_i \bowtie \langle \mathbf{Z} \rangle \mathbf{W}_i \left[(\mathbf{A}\boldsymbol{\Sigma}_{\boldsymbol{x}_i} - \mathbf{A}\boldsymbol{\Sigma}_{\boldsymbol{x}_i} \mathbf{A}^{\mathrm{T}} \mathbf{W}_i \boldsymbol{\Xi}_i) \right. \\
&\qquad \left. - (\boldsymbol{\Sigma}_{\boldsymbol{x}_i \boldsymbol{y}_i} - \boldsymbol{\Sigma}_{\boldsymbol{x}_i \boldsymbol{y}_i} \mathbf{A}^{\mathrm{T}} \mathbf{W}_i \boldsymbol{\Xi}_i) + \mathbf{A}\boldsymbol{\Sigma}_{\boldsymbol{x}_i \boldsymbol{y}_i}^{\mathrm{T}} \mathbf{W}_i \boldsymbol{\Xi}_i - \boldsymbol{\Sigma}_{\boldsymbol{y}_i} \mathbf{W}_i \boldsymbol{\Xi}_i \right] \\
&= \quad \mathbf{W}_i \, \boldsymbol{r}_i \bowtie \langle \mathbf{Z} \rangle \mathbf{W}_i \left[(\mathbf{A}\boldsymbol{\Sigma}_{\boldsymbol{x}_i} \mathbf{A}^{\mathrm{T}} - \boldsymbol{\Sigma}_{\boldsymbol{x}_i \boldsymbol{y}_i} \mathbf{A}^{\mathrm{T}} - \mathbf{A}\boldsymbol{\Sigma}_{\boldsymbol{x}_i \boldsymbol{y}_i}^{\mathrm{T}} + \boldsymbol{\Sigma}_{\boldsymbol{y}_i}) \mathbf{W}_i \boldsymbol{\Xi}_i - (\mathbf{A}\boldsymbol{\Sigma}_{\boldsymbol{x}_i} - \boldsymbol{\Sigma}_{\boldsymbol{x}_i \boldsymbol{y}_i}) \right] \\
&= \quad \mathbf{W}_i \, \boldsymbol{r}_i \bowtie \langle \mathbf{Z} \rangle \left[\mathbf{W}_i \boldsymbol{\Xi}_i - \mathbf{W}_i \boldsymbol{\Xi}_i \right] \; = \; \mathbf{0} \\
[8] \quad &= \quad \mathbf{0} \\
[9] \quad &= \quad \langle \mathbf{Z} \rangle \, \mathbf{W}_i \left[\mathbf{A}\boldsymbol{\Sigma}_{\boldsymbol{x}_i} \mathbf{A}^{\mathrm{T}} - \mathbf{A}\boldsymbol{\Sigma}_{\boldsymbol{x}_i \boldsymbol{y}_i}^{\mathrm{T}} - \boldsymbol{\Sigma}_{\boldsymbol{x}_i \boldsymbol{y}_i} \mathbf{A}^{\mathrm{T}} + \boldsymbol{\Sigma}_{\boldsymbol{y}_i} \right] \mathbf{W}_i \langle \mathbf{Z} \rangle^{\mathrm{T}} \\
&= \quad \langle \mathbf{Z} \rangle \, \mathbf{W}_i \langle \mathbf{Z} \rangle^{\mathrm{T}}
\end{aligned}
$$

Thus removing hats from (72) and adding the new terms, we get

$$
\begin{aligned}
\mathbf{Q} \quad = \quad & \sum_i \left\{ \boldsymbol{\mu}_i \boldsymbol{\mu}_i^{\mathrm{T}} \otimes \mathbf{W}_i + \boldsymbol{\Sigma}_{\boldsymbol{x}_i} \otimes \mathbf{W}_i \boldsymbol{r}_i \boldsymbol{r}_i^{\mathrm{T}} \mathbf{W}_i - \boldsymbol{\Xi}_i^{\mathrm{T}} \mathbf{W}_i \boldsymbol{\Xi}_i \otimes \mathbf{W}_i \boldsymbol{r}_i \boldsymbol{r}_i^{\mathrm{T}} \mathbf{W}_i \right\} \\
& - \left(\sum_i \boldsymbol{\mu}_i \otimes \mathbf{W}_i \right) \langle \mathbf{Z} \rangle^{\mathrm{T}} - \langle \mathbf{Z} \rangle \left(\sum_i \boldsymbol{\mu}_i^{\mathrm{T}} \otimes \mathbf{W}_i \right) \\
& + \langle \mathbf{Z} \rangle \left(\sum_i \mathbf{W}_i \right) \langle \mathbf{Z} \rangle^{\mathrm{T}}.
\end{aligned}
\tag{99}
$$

Using $\sum_i \mathbf{W}_i = \boldsymbol{\Omega}$ and the definition of $\langle \mathbf{Z} \rangle$ in (97), the last terms are

$$
\begin{aligned}
& -\langle \mathbf{M} \rangle \, \boldsymbol{\Omega} \, (\langle \mathbf{M} \rangle - \langle \mathbf{E} \rangle)^{\mathrm{T}} - (\langle \mathbf{M} \rangle - \langle \mathbf{E} \rangle) \, \boldsymbol{\Omega} \, \langle \mathbf{M} \rangle^{\mathrm{T}} + (\langle \mathbf{M} \rangle - \langle \mathbf{E} \rangle) \, \boldsymbol{\Omega} \, (\langle \mathbf{M} \rangle - \langle \mathbf{E} \rangle)^{\mathrm{T}} \\
&= -\langle \mathbf{M} \rangle \, \boldsymbol{\Omega} \langle \mathbf{M} \rangle^{\mathrm{T}} + \langle \mathbf{M} \rangle \, \boldsymbol{\Omega} \langle \mathbf{E} \rangle^{\mathrm{T}} - \langle \mathbf{M} \rangle \, \boldsymbol{\Omega} \langle \mathbf{M} \rangle^{\mathrm{T}} + \langle \mathbf{E} \rangle \, \boldsymbol{\Omega} \langle \mathbf{M} \rangle^{\mathrm{T}} \\
&\qquad + \langle \mathbf{M} \rangle \, \boldsymbol{\Omega} \langle \mathbf{M} \rangle^{\mathrm{T}} - \langle \mathbf{M} \rangle \, \boldsymbol{\Omega} \langle \mathbf{E} \rangle^{\mathrm{T}} - \langle \mathbf{E} \rangle \, \boldsymbol{\Omega} \langle \mathbf{M} \rangle^{\mathrm{T}} + \langle \mathbf{E} \rangle \, \boldsymbol{\Omega} \langle \mathbf{E} \rangle^{\mathrm{T}} \\
&= -\left(\langle \mathbf{M} \rangle \, \boldsymbol{\Omega} \langle \mathbf{M} \rangle^{\mathrm{T}} - \langle \mathbf{E} \rangle \, \boldsymbol{\Omega} \langle \mathbf{E} \rangle^{\mathrm{T}} \right)
\end{aligned}
$$

Therefore

$$
\begin{aligned}
\mathbf{Q} \quad = \quad & \sum_i \left\{ \boldsymbol{\mu}_i \boldsymbol{\mu}_i^{\mathrm{T}} \otimes \mathbf{W}_i + \boldsymbol{\Sigma}_{\boldsymbol{x}_i} \otimes \mathbf{W}_i \boldsymbol{r}_i \boldsymbol{r}_i^{\mathrm{T}} \mathbf{W}_i - \boldsymbol{\Xi}_i^{\mathrm{T}} \mathbf{W}_i \boldsymbol{\Xi}_i \otimes \mathbf{W}_i \boldsymbol{r}_i \boldsymbol{r}_i^{\mathrm{T}} \mathbf{W}_i \right\} \\
& - \left(\langle \mathbf{M} \rangle \, \boldsymbol{\Omega} \langle \mathbf{M} \rangle^{\mathrm{T}} - \langle \mathbf{E} \rangle \, \boldsymbol{\Omega} \langle \mathbf{E} \rangle^{\mathrm{T}} \right).
\end{aligned}
\tag{100}
$$

Now we turn to the calculation of $\boldsymbol{\Sigma}_{\boldsymbol{a}_0}$. From (93),

$$
\boldsymbol{\Sigma}_{\boldsymbol{a}_0} \quad = \quad \sum_i \frac{\mathrm{d}\boldsymbol{a}_0}{\mathrm{d}\mathbb{z}_i} \Sigma_i \left[\frac{\mathrm{d}\boldsymbol{a}_0}{\mathrm{d}\mathbb{z}_i} \right]^{\mathrm{T}}
$$

$$
= \sum_i \left[\frac{\partial \boldsymbol{a}_0}{\partial \boldsymbol{z}_i} + \frac{\partial \boldsymbol{a}_0}{\partial \mathbf{A}} \frac{\mathrm{d}\mathbf{A}}{\mathrm{d}\boldsymbol{z}_i} \right] \Sigma_i \left[\frac{\partial \boldsymbol{a}_0}{\partial \boldsymbol{z}_i} + \frac{\partial \boldsymbol{a}_0}{\partial \mathbf{A}} \frac{\mathrm{d}\mathbf{A}}{\mathrm{d}\boldsymbol{z}_i} \right]^{\mathrm{T}}
$$

$$
= \sum_i \left\{ \frac{\partial \boldsymbol{a}_0}{\partial \boldsymbol{z}_i} \Sigma_i \left[\frac{\partial \boldsymbol{a}_0}{\partial \boldsymbol{z}_i} \right]^{\mathrm{T}} + \frac{\partial \boldsymbol{a}_0}{\partial \mathbf{A}} \frac{\mathrm{d}\mathbf{A}}{\mathrm{d}\boldsymbol{z}_i} \Sigma_i \left[\frac{\mathrm{d}\mathbf{A}}{\mathrm{d}\boldsymbol{z}_i} \right]^{\mathrm{T}} \left[\frac{\partial \boldsymbol{a}_0}{\partial \mathbf{A}} \right]^{\mathrm{T}} \right.
$$

$$
\left. + \frac{\partial \boldsymbol{a}_0}{\partial \boldsymbol{z}_i} \Sigma_i \left[\frac{\mathrm{d}\mathbf{A}}{\mathrm{d}\boldsymbol{z}_i} \right]^{\mathrm{T}} \left[\frac{\partial \boldsymbol{a}_0}{\partial \mathbf{A}} \right]^{\mathrm{T}} + \frac{\partial \boldsymbol{a}_0}{\partial \mathbf{A}} \frac{\mathrm{d}\mathbf{A}}{\mathrm{d}\boldsymbol{z}_i} \Sigma_i \left[\frac{\partial \boldsymbol{a}_0}{\partial \boldsymbol{z}_i} \right]^{\mathrm{T}} \right\}. \tag{101}
$$

Differentiating (44) we obtain

$$
\frac{\partial \boldsymbol{a}_0}{\partial \boldsymbol{x}_i} = -\boldsymbol{\Omega}^{-1} \mathbf{W}_i \mathbf{A}, \quad i = 1, \dots n \tag{102}
$$

$$
\frac{\partial \boldsymbol{a}_0}{\partial \boldsymbol{y}_i} = \boldsymbol{\Omega}^{-1} \mathbf{W}_i, \quad i = 1, \dots n \tag{103}
$$

So the first term in (101) is

$$
\sum_i \frac{\partial \boldsymbol{a}_0}{\partial \boldsymbol{z}_i} \Sigma_i \left[\frac{\partial \boldsymbol{a}_0}{\partial \boldsymbol{z}_i} \right]^{\mathrm{T}} = \sum_i \boldsymbol{\Omega}^{-1} \mathbf{W}_i \begin{bmatrix} -\mathbf{A} & \mathbf{I}_{n_y} \end{bmatrix} \begin{bmatrix} \boldsymbol{\Sigma}_{\boldsymbol{x}_i} & \boldsymbol{\Sigma}_{\boldsymbol{x}_i \boldsymbol{y}_i}^{\mathrm{T}} \\ \boldsymbol{\Sigma}_{\boldsymbol{x}_i \boldsymbol{y}_i} & \boldsymbol{\Sigma}_{\boldsymbol{y}_i} \end{bmatrix} \begin{bmatrix} -\mathbf{A}^{\mathrm{T}} \\ \mathbf{I}_{n_y} \end{bmatrix} \mathbf{W}_i \boldsymbol{\Omega}^{-1}
$$

$$
= \boldsymbol{\Omega}^{-1} \sum_i \mathbf{W}_i \left[\mathbf{A} \boldsymbol{\Sigma}_{\boldsymbol{x}_i} \mathbf{A}^{\mathrm{T}} - \mathbf{A} \boldsymbol{\Sigma}_{\boldsymbol{x}_i \boldsymbol{y}_i}^{\mathrm{T}} - \boldsymbol{\Sigma}_{\boldsymbol{x}_i \boldsymbol{y}_i} \mathbf{A}^{\mathrm{T}} + \boldsymbol{\Sigma}_{\boldsymbol{y}_i} \right] \mathbf{W}_i \boldsymbol{\Omega}^{-1}
$$

$$
= \boldsymbol{\Omega}^{-1} \sum_i \mathbf{W}_i \boldsymbol{\Omega}^{-1} = \boldsymbol{\Omega}^{-1}.
$$

Now, $\partial \boldsymbol{a}_0 / \partial \mathbf{A}$ is already worked out on pages 19 ff. as the second term of (76), and appears as the second term in (80). Using the definitions introduced subsequently to that point, it becomes

$$
\frac{\partial \boldsymbol{a}_0}{\partial \mathbf{A}} = \boldsymbol{\Omega}^{-1} \sum_{i=1}^{m} \left\{ \mathbf{W}_i \boldsymbol{\Xi}_i \bowtie \mathbf{W}_i \boldsymbol{r}_i - \boldsymbol{\mu}_i^{\mathrm{T}} \otimes \mathbf{W}_i \right\}
$$

$$
= \left(\langle \mathbf{E} \rangle - \langle \mathbf{M} \rangle \right)^{\mathrm{T}} = -\langle \mathbf{Z} \rangle^{\mathrm{T}}. \tag{104}
$$

So the second term in (101) is

$$
\sum_i \frac{\partial \boldsymbol{a}_0}{\partial \mathbf{A}} \frac{\mathrm{d}\mathbf{A}}{\mathrm{d}\boldsymbol{z}_i} \Sigma_i \left[\frac{\mathrm{d}\mathbf{A}}{\mathrm{d}\boldsymbol{z}_i} \right]^{\mathrm{T}} \left[\frac{\partial \boldsymbol{a}_0}{\partial \mathbf{A}} \right]^{\mathrm{T}} = \frac{\partial \boldsymbol{a}_0}{\partial \mathbf{A}} \boldsymbol{\Sigma}_{\mathbf{A}} \left[\frac{\partial \boldsymbol{a}_0}{\partial \mathbf{A}} \right]^{\mathrm{T}}
$$

$$
= \langle \mathbf{Z} \rangle^{\mathrm{T}} \boldsymbol{\Sigma}_{\mathbf{A}} \langle \mathbf{Z} \rangle.
$$

The last two terms in (101) are transposes of each other. Using (39),

$$
\frac{\mathrm{d}\mathbf{A}}{\mathrm{d}\boldsymbol{z}_i} = -\mathbf{J}^{-1} \frac{\partial \mathbf{F}}{\partial \boldsymbol{z}_i} = -\mathbf{J}^{-1} \begin{bmatrix} \mathbf{F}_{\boldsymbol{x}_i}, & \mathbf{F}_{\boldsymbol{y}_i} \end{bmatrix},
$$

the first is

$$
\sum_i \frac{\partial \boldsymbol{a}_0}{\partial \boldsymbol{z}_i} \Sigma_i \left[\frac{\mathrm{d}\mathbf{A}}{\mathrm{d}\boldsymbol{z}_i} \right]^{\mathrm{T}} \left[\frac{\partial \boldsymbol{a}_0}{\partial \mathbf{A}} \right]^{\mathrm{T}}
$$

$$
= \sum_i \boldsymbol{\Omega}^{-1} \mathbf{W}_i \begin{bmatrix} -\mathbf{A} & \mathbf{I}_{n_y} \end{bmatrix} \begin{bmatrix} \boldsymbol{\Sigma}_{\boldsymbol{x}_i} & \boldsymbol{\Sigma}_{\boldsymbol{x}_i \boldsymbol{y}_i}^{\mathrm{T}} \\ \boldsymbol{\Sigma}_{\boldsymbol{x}_i \boldsymbol{y}_i} & \boldsymbol{\Sigma}_{\boldsymbol{y}_i} \end{bmatrix} \begin{bmatrix} \mathbf{F}_{\boldsymbol{x}_i}^{\mathrm{T}} \\ \mathbf{F}_{\boldsymbol{y}_i}^{\mathrm{T}} \end{bmatrix} \mathbf{J}^{-1} \langle \mathbf{Z} \rangle
$$

$$
= \boldsymbol{\Omega}^{-1} \sum_i \mathbf{W}_i \left[-\mathbf{A} \boldsymbol{\Sigma}_{\boldsymbol{x}_i} \mathbf{F}_{\boldsymbol{x}_i}^{\mathrm{T}} - \mathbf{A} \boldsymbol{\Sigma}_{\boldsymbol{x}_i \boldsymbol{y}_i}^{\mathrm{T}} \mathbf{F}_{\boldsymbol{y}_i}^{\mathrm{T}} + \boldsymbol{\Sigma}_{\boldsymbol{x}_i \boldsymbol{y}_i} \mathbf{F}_{\boldsymbol{x}_i}^{\mathrm{T}} + \boldsymbol{\Sigma}_{\boldsymbol{y}_i} \mathbf{F}_{\boldsymbol{y}_i}^{\mathrm{T}} \right] \mathbf{J}^{-1} \langle \mathbf{Z} \rangle
$$

The terms inside brackets are

$$-\mathbf{A}\boldsymbol{\Sigma}_{\boldsymbol{x}_i}\mathbf{F}_{\boldsymbol{x}_i}^{\mathrm{T}}$$
$$= -\mathbf{A}\boldsymbol{\Sigma}_{\boldsymbol{x}_i}\left[\boldsymbol{\mu}_i^{\mathrm{T}}\otimes\mathbf{A}^{\mathrm{T}}\mathbf{W}_i + (\mathbf{I}_{n_x} - \mathbf{A}^{\mathrm{T}}\mathbf{W}_i\boldsymbol{\Xi}_i)\bowtie\mathbf{W}_i\,\boldsymbol{r}_i - \mathbf{A}^{\mathrm{T}}\mathbf{W}_i\langle\mathbf{Z}\rangle^{\mathrm{T}}\right]$$
$$= -\boldsymbol{\mu}_i^{\mathrm{T}}\otimes\mathbf{A}\boldsymbol{\Sigma}_{\boldsymbol{x}_i}\mathbf{A}^{\mathrm{T}}\mathbf{W}_i - \mathbf{A}\boldsymbol{\Sigma}_{\boldsymbol{x}_i}\bowtie\mathbf{W}_i\,\boldsymbol{r}_i + \mathbf{A}\boldsymbol{\Sigma}_{\boldsymbol{x}_i}\mathbf{A}^{\mathrm{T}}\mathbf{W}_i\boldsymbol{\Xi}_i\bowtie\mathbf{W}_i\,\boldsymbol{r}_i + \mathbf{A}\boldsymbol{\Sigma}_{\boldsymbol{x}_i}\mathbf{A}^{\mathrm{T}}\mathbf{W}_i\langle\mathbf{Z}\rangle^{\mathrm{T}}$$

$$-\mathbf{A}\boldsymbol{\Sigma}_{\boldsymbol{x}_i\boldsymbol{y}_i}^{\mathrm{T}}\mathbf{F}_{\boldsymbol{y}_i}^{\mathrm{T}}$$
$$= -\mathbf{A}\boldsymbol{\Sigma}_{\boldsymbol{x}_i\boldsymbol{y}_i}^{\mathrm{T}}\left[-\boldsymbol{\mu}_i^{\mathrm{T}}\otimes\mathbf{W}_i + \mathbf{W}_i\boldsymbol{\Xi}_i\bowtie\mathbf{W}_i\,\boldsymbol{r}_i + \mathbf{W}_i\langle\mathbf{Z}\rangle^{\mathrm{T}}\right]$$
$$= \boldsymbol{\mu}_i^{\mathrm{T}}\otimes\mathbf{A}\boldsymbol{\Sigma}_{\boldsymbol{x}_i\boldsymbol{y}_i}^{\mathrm{T}}\mathbf{W}_i - \mathbf{A}\boldsymbol{\Sigma}_{\boldsymbol{x}_i\boldsymbol{y}_i}^{\mathrm{T}}\mathbf{W}_i\boldsymbol{\Xi}_i\bowtie\mathbf{W}_i\,\boldsymbol{r}_i - \mathbf{A}\boldsymbol{\Sigma}_{\boldsymbol{x}_i\boldsymbol{y}_i}^{\mathrm{T}}\mathbf{W}_i\langle\mathbf{Z}\rangle^{\mathrm{T}}$$

$$\boldsymbol{\Sigma}_{\boldsymbol{x}_i\boldsymbol{y}_i}\mathbf{F}_{\boldsymbol{x}_i}^{\mathrm{T}}$$
$$= \boldsymbol{\Sigma}_{\boldsymbol{x}_i\boldsymbol{y}_i}\left[\boldsymbol{\mu}_i^{\mathrm{T}}\otimes\mathbf{A}^{\mathrm{T}}\mathbf{W}_i + (\mathbf{I}_{n_x} - \mathbf{A}^{\mathrm{T}}\mathbf{W}_i\boldsymbol{\Xi}_i)\bowtie\mathbf{W}_i\,\boldsymbol{r}_i - \mathbf{A}^{\mathrm{T}}\mathbf{W}_i\langle\mathbf{Z}\rangle^{\mathrm{T}}\right]$$
$$= \boldsymbol{\mu}_i^{\mathrm{T}}\otimes\boldsymbol{\Sigma}_{\boldsymbol{x}_i\boldsymbol{y}_i}\mathbf{A}^{\mathrm{T}}\mathbf{W}_i + \boldsymbol{\Sigma}_{\boldsymbol{x}_i\boldsymbol{y}_i}\bowtie\mathbf{W}_i\,\boldsymbol{r}_i - \boldsymbol{\Sigma}_{\boldsymbol{x}_i\boldsymbol{y}_i}\mathbf{A}^{\mathrm{T}}\mathbf{W}_i\boldsymbol{\Xi}_i\bowtie\mathbf{W}_i\,\boldsymbol{r}_i - \boldsymbol{\Sigma}_{\boldsymbol{x}_i\boldsymbol{y}_i}\mathbf{A}^{\mathrm{T}}\mathbf{W}_i\langle\mathbf{Z}\rangle^{\mathrm{T}}$$

$$\boldsymbol{\Sigma}_{\boldsymbol{y}_i}\mathbf{F}_{\boldsymbol{y}_i}^{\mathrm{T}}$$
$$= \boldsymbol{\Sigma}_{\boldsymbol{y}_i}\left[-\boldsymbol{\mu}_i^{\mathrm{T}}\otimes\mathbf{W}_i + \mathbf{W}_i\boldsymbol{\Xi}_i\bowtie\mathbf{W}_i\,\boldsymbol{r}_i + \mathbf{W}_i\langle\mathbf{Z}\rangle^{\mathrm{T}}\right]$$
$$= -\boldsymbol{\mu}_i^{\mathrm{T}}\otimes\boldsymbol{\Sigma}_{\boldsymbol{y}_i}\mathbf{W}_i + \boldsymbol{\Sigma}_{\boldsymbol{y}_i}\mathbf{W}_i\boldsymbol{\Xi}_i\bowtie\mathbf{W}_i\,\boldsymbol{r}_i + \boldsymbol{\Sigma}_{\boldsymbol{y}_i}\mathbf{W}_i\langle\mathbf{Z}\rangle^{\mathrm{T}}$$

Adding these together yields

$$\boldsymbol{\mu}_i^{\mathrm{T}}\otimes\left[(-\mathbf{A}\boldsymbol{\Sigma}_{\boldsymbol{x}_i}\mathbf{A}^{\mathrm{T}} + \mathbf{A}\boldsymbol{\Sigma}_{\boldsymbol{x}_i\boldsymbol{y}_i}^{\mathrm{T}} + \boldsymbol{\Sigma}_{\boldsymbol{x}_i\boldsymbol{y}_i}\mathbf{A}^{\mathrm{T}} - \boldsymbol{\Sigma}_{\boldsymbol{y}_i})\right]\mathbf{W}_i$$
$$-(\mathbf{A}\boldsymbol{\Sigma}_{\boldsymbol{x}_i} - \boldsymbol{\Sigma}_{\boldsymbol{x}_i\boldsymbol{y}_i})\bowtie\mathbf{W}_i\,\boldsymbol{r}_i$$
$$+\left[(\mathbf{A}\boldsymbol{\Sigma}_{\boldsymbol{x}_i}\mathbf{A}^{\mathrm{T}} - \mathbf{A}\boldsymbol{\Sigma}_{\boldsymbol{x}_i\boldsymbol{y}_i}^{\mathrm{T}} - \boldsymbol{\Sigma}_{\boldsymbol{x}_i\boldsymbol{y}_i}\mathbf{A}^{\mathrm{T}} + \boldsymbol{\Sigma}_{\boldsymbol{y}_i})\mathbf{W}_i\boldsymbol{\Xi}_i\right]\bowtie\mathbf{W}_i\,\boldsymbol{r}_i$$
$$+\left[\mathbf{A}\boldsymbol{\Sigma}_{\boldsymbol{x}_i}\mathbf{A}^{\mathrm{T}} - \mathbf{A}\boldsymbol{\Sigma}_{\boldsymbol{x}_i\boldsymbol{y}_i}^{\mathrm{T}} - \boldsymbol{\Sigma}_{\boldsymbol{x}_i\boldsymbol{y}_i}\mathbf{A}^{\mathrm{T}} + \boldsymbol{\Sigma}_{\boldsymbol{y}_i}\right]\mathbf{W}_i\langle\mathbf{Z}\rangle^{\mathrm{T}}$$
$$= -\boldsymbol{\mu}_i^{\mathrm{T}}\otimes\mathbf{I}_{n_y} - \boldsymbol{\Xi}_i\bowtie\mathbf{W}_i\,\boldsymbol{r}_i + \boldsymbol{\Xi}_i\bowtie\mathbf{W}_i\,\boldsymbol{r}_i + \langle\mathbf{Z}\rangle^{\mathrm{T}}$$
$$= -\mathbf{M}_i^{\mathrm{T}} + \langle\mathbf{Z}\rangle^{\mathrm{T}}$$

Summation then gives for this whole term

$$\boldsymbol{\Omega}^{-1}\sum_i\mathbf{W}_i\left[-\mathbf{M}_i + \langle\mathbf{Z}\rangle\right]^{\mathrm{T}}\mathbf{J}^{-1}\langle\mathbf{Z}\rangle = (-\langle\mathbf{M}\rangle + \langle\mathbf{Z}\rangle)^{\mathrm{T}}\mathbf{J}^{-1}\langle\mathbf{Z}\rangle = -\langle\mathbf{E}\rangle^{\mathrm{T}}\mathbf{J}^{-1}\langle\mathbf{Z}\rangle$$

The other term is the transpose of this one. So we finally obtain

$$\boldsymbol{\Sigma}_{\boldsymbol{a}_0} = \boldsymbol{\Omega}^{-1} + \langle\mathbf{Z}\rangle^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathbf{A}}\langle\mathbf{Z}\rangle - \langle\mathbf{E}\rangle^{\mathrm{T}}\mathbf{J}^{-1}\langle\mathbf{Z}\rangle - \langle\mathbf{Z}\rangle^{\mathrm{T}}\mathbf{J}^{-1}\langle\mathbf{E}\rangle \tag{105}$$

Finally, we solve for $\boldsymbol{\Sigma}_{\boldsymbol{a}_0,\mathbf{A}}$, the $n_x n_y \times n_y$ matrix in the lower left of $\boldsymbol{\Sigma}_{\hat{\mathbf{A}}}$. From (93),

$$\boldsymbol{\Sigma}_{\boldsymbol{a}_0,\mathbf{A}} = \sum_i\frac{\mathrm{d}\mathbf{A}}{\mathrm{d}\boldsymbol{\varkappa}_i}\boldsymbol{\Sigma}_i\left[\frac{\mathrm{d}\boldsymbol{a}_0}{\mathrm{d}\boldsymbol{\varkappa}_i}\right]^{\mathrm{T}}$$
$$= \sum_i\left[\frac{\mathrm{d}\mathbf{A}}{\mathrm{d}\boldsymbol{\varkappa}_i}\right]\boldsymbol{\Sigma}_i\left[\frac{\partial\boldsymbol{a}_0}{\partial\boldsymbol{\varkappa}} + \frac{\partial\boldsymbol{a}_0}{\partial\mathbf{A}}\frac{\mathrm{d}\mathbf{A}}{\mathrm{d}\boldsymbol{\varkappa}_i}\right]^{\mathrm{T}}$$
$$= \sum_i\frac{\partial\boldsymbol{a}_0}{\partial\boldsymbol{\varkappa}}\boldsymbol{\Sigma}_i\left[\frac{\mathrm{d}\mathbf{A}}{\mathrm{d}\boldsymbol{\varkappa}_i}\right]^{\mathrm{T}} + \boldsymbol{\Sigma}_{\mathbf{A}}\left[\frac{\partial\boldsymbol{a}_0}{\partial\mathbf{A}}\right]^{\mathrm{T}}. \tag{106}$$

The first part of this expression is the same as the coefficient of $(\partial\boldsymbol{a}_0/\partial\mathbf{A})$ in the fourth term in (101). The transpose of this coefficient was worked out in obtaining (105), giving here $-\mathbf{J}^{-1}\langle\mathbf{E}\rangle$. Using also the result for $(\partial\boldsymbol{a}_0/\partial\mathbf{A})$ in (104) we have

$$\boldsymbol{\Sigma}_{\boldsymbol{a}_0,\mathbf{A}} = \mathbf{J}^{-1}\langle\mathbf{E}\rangle - \boldsymbol{\Sigma}_{\mathbf{A}}\langle\mathbf{Z}\rangle. \tag{107}$$

References

[1] R. Baldick, Applied Optimization: Formulation And Algorithms For Engineering Systems, Cambridge University Press, New York, 2005.

[2] P.R. Bevington and D.K. Robinson, Data Reduction and Error Analysis for the Physical Sciences, 3rd ed., McGraw-Hill, New York, 2003.

[3] W.E. Deming, Statistical Adjustment of Data, Wiley, New York, 1943.

[4] G.H. Golub and C.F. Van Loan, Matrix Computations, Johns Hopkins U. Press, Baltimore, Maryland, 1983.

[5] D. Gubbins, A.L. Jones, and C.C. Finlay, Fall in Earth's Magnetic Field is Erratic, Science 312 (2006) 900–902.

[6] J.T. Kent, G.S. Watson, and T.C. Onstott, Fitting straight lines and planes with an application to radiometric dating, Earth Planet. Sci. Lett. 97 (1990) 1–17.

[7] J.R. Magnus, private communication (2006).

[8] J.R. Magnus and H. Neudecker, Matrix Differential Calculus with Applications in Statistics and Econometrics, Wiley, New York, 1988.

[9] J.H. Reynolds, E.C. Alexander, Jr., P.K. Davis, and B. Srinivasan, Studies of K-Ar dating and xenon from extinct radioactivities in breccia 14318; implications for early lunar history, Geochim. Cosmochim. Acta 38 (1974) 401–417.

[10] J.H. Williamson, Least-squares fitting of a straight line, Can. J. Phys. 46 (1968) 1845–1846.

[11] D. York, Least-squares fitting of a straight line, Can. J. Phys. 44 (1966) 1079–1086.

[12] D. York, Least-squares fitting of a straight line with correlated errors, Earth Planet. Sci. Lett. 5 (1969) 320–324.