2015

Figures of merit for simultaneous inference and comparisons in simulation experiments

Noel Cressie
University of Wollongong

Sandy Burden
University of Wollongong

Figures of merit for simultaneous inference and comparisons in simulation experiments

Abstract

This article considers the traditional figures of merit (FOMs), namely bias, and mean squared (prediction) error, that are typically used to evaluate simulation experiments. We propose functions of them that account for different variables' units; these alternative FOMs are closely tied to simultaneous multivariate inference on an unknown parameter vector or unknown state vector. Their usefulness is illustrated in a simulation experiment, where the goal is to determine the statistical properties associated with prediction of a multivariate state.

NIASRA

NATIONAL INSTITUTE FOR APPLIED STATISTICS RESEARCH AUSTRALIA

National Institute for Applied Statistics Research Australia

The University of Wollongong

Working Paper

12-15

Figures of Merit for Simultaneous Inference and Comparisons in Simulation Experiments

Noel Cressie and Sandy Burden

National Institute for Applied Statistics Research Australia, University of Wollongong,
Wollongong NSW 2522. Phone +61 2 4221 5435, Fax +61 2 4221 4845.
Email: anica@uow.edu.au

Figures of Merit for Simultaneous Inference and Comparisons in Simulation Experiments

Noel Cressie[a,b]*, Sandy Burden[a]

This article considers the traditional figures of merit (FOMs), namely bias, and mean squared (prediction) error, that are typically used to evaluate simulation experiments. We propose functions of them that account for different variables' units; these alternative FOMs are closely tied to simultaneous multivariate inference on an unknown parameter vector or unknown state vector. Their usefulness is illustrated in a simulation experiment, where the goal is to determine the statistical properties associated with prediction of a multivariate state.

Keywords: bias; correlation matrix; inverse coefficient of variation; mean squared prediction error; prediction interval; prediction region; standard deviation; Wald statistic

1. Introduction

Simulation experiments are designed to determine statistical properties associated with estimation of a fixed parameter θ or prediction of a random state x. In the first scenario, data y are generated from $f_\theta(y)$, from which an estimator of θ, $\hat{\theta}(y)$, is computed. The statistical properties of $\hat{\theta}(y)$ come from those of y, and often simulation is the only way to obtain them. The two traditional figures of merit (FOMs) used to evaluate $\hat{\theta}(y)$ are:

$$Bias \equiv E_y(\hat{\theta}(y)) - \theta \; ; \qquad Mse \equiv E_y(\hat{\theta}(y) - \theta)^2,$$

namely the *bias* and the *mean squared error*, respectively. The subscript 'y' in 'E_y' signifies that the expectation is taken with respect to the random quantity y.

In the second scenario, a state x is generated from a probability distribution $g(x)$ and, conditional on x, the data y are generated from $f(y|x)$; the predictor of x, $\hat{x}(y)$, is then computed. Because the state x is unobserved, the statistical properties associated with the predictor come from the joint distribution of y and x, namely $f(y|x)g(x)$. That is, they come from the statistical properties of x and from the statistical properties of y conditional on x. The prediction error is defined to be,

$$\hat{x}(y) - x, \tag{1}$$

[a]**Centre for Environmental Informatics, NIASRA, University of Wollongong, Wollongong, NSW 2522, Australia**
[b]**Distinguished Visiting Scientist, Jet Propulsion Laboratory, Pasadena, CA 91109**
*__Email: ncressie@uow.edu.au__

and the traditional FOMs are based on its first two moments. Specifically, they are

$$Bias \equiv E_{y,x}(\hat{x}(y) - x) \; ; \qquad Mspe \equiv E_{y,x}(\hat{x}(y) - x)^2,$$

which are called the *bias* and the *mean squared prediction error*, respectively. Notice that $Bias = E_y(\hat{x}(y)) - E_x(x)$, after marginalisation, and subscripts on the expectation operator are maintained for clarity.

In what follows, we shall concentrate on the second scenario, where we wish to evaluate predictors $\hat{x}(y)$ of a random state x. The initial scenario described, of parameter estimation, has an analogous treatment. Indeed, there is a more general scenario where both the random state x *and* a fixed parameter θ are unknown. While this adds complexity to the statistical methodology used to derive $\hat{x}(y)$, it does not change what follows, since a predictor of x has to be a statistic, namely a function only of the data y. These simulation-experiment scenarios are different from that of a "computer experiment." There, a computationally expensive, non-stochastic algorithm is typically run for certain chosen factor-level combinations. Inference is then made on the responses at factor-level combinations where computations were not performed; see, for example, Santner et al. (2003) for further details. In contrast, a simulation experiment is run many times at each of the chosen factor-level combinations, and the prediction properties of $\hat{x}(y)$ are compared (using FOMs) for these factor-level combinations.

Although we do not explicitly use "bold" notation, we consider the state x and its predictor $\hat{x}(y)$ to be possibly multivariate of dimension $K = \{1, 2, ...\}$, and hence *Bias* could be a K-dimensional vector and *Mspe* could be a $K \times K$ matrix:

$$Bias \equiv E_{y,x}(\hat{x}(y) - x) \; ; \qquad Mspe \equiv E_{y,x}\left((\hat{x}(y) - x)(\hat{x}(y) - x)'\right). \tag{2}$$

When needed, we write, $Bias = (Bias_1, ..., Bias_K)'$.

We shall see in Section 3.2 that, when making simultaneous inference on the elements of x, all of which may have completely different units (e.g., parts per million by volume, hectopascals, and degrees Celsius), other FOMs are more natural. First define

$$Cov \equiv Mspe - (Bias)(Bias)', \tag{3}$$

which represents the covariance matrix, $\mathrm{cov}_{y,x}(\hat{x}(y) - x)$, henceforth assumed to be positive-definite. When needed we write $Cov = (Cov_{kk'})$. Next define the diagonal matrix,

$$Var \equiv \mathrm{diag}(Cov), \tag{4}$$

whose non-zero elements run down the diagonal and represent the variances of the elements of $\hat{x}(y) - x$. Since (3) is positive-definite, then all variances are strictly positive, and hence the diagonal matrix, $Var^{-1/2}$, is well defined. Then the correlation matrix associated with the prediction error, $\hat{x}(y) - x$, is

$$Var^{-1/2}(Cov)Var^{-1/2}, \tag{5}$$

which is positive-definite.

Our goal in defining alternative FOMs to those given in (2) is to account for different units transparently and to allow for meaningful comparisons in simulation experiments. The FOM,

$$Icv \equiv (Var)^{-1/2}Bias, \tag{6}$$

is a K-dimensional vector of unit-free elements of bias, whose k-th element is given by $Bias_k/(Cov_{kk})^{1/2}$; $k = 1, ..., K$. This is the inverse of the coefficient of variation, and hence we use the notation *Icv* in (6).

The FOM,

$$Sdv \equiv (Var)^{1/2}, \tag{7}$$

is a $K \times K$ diagonal matrix, whose elements on the diagonal are standard deviations with units that are respectively the units of the K-dimensional state x. Then the FOM,

$$Cor \equiv Sdv^{-1}(Cov)Sdv^{-1}, \tag{8}$$

which is equivalently given by (5), is a $K \times K$ correlation matrix of unit-free elements all of whose diagonal elements are 1 and whose off-diagonal elements are between -1 and 1, capturing the statistical dependence between elements of the prediction error. Clearly, we can write $Cor = \left(\frac{Cov_{kk'}}{(Cov_{kk})^{1/2}(Cov_{k'k'})^{1/2}}\right)$, and recall that Cov and hence Cor are positive-definite.

The FOM (8) can be represented spectrally as,

$$Cor = Eig(Lam)Eig,$$

where Lam is a diagonal matrix of K positive, ordered (from largest to smallest) eigenvalues and Eig is a $K \times K$ matrix whose columns are orthonormal eigenvectors for which $(Eig)'Eig = Eig(Eig)' = I$, the identity matrix of order K. Let Lam_k be the eigenvalue corresponding to the eigenvector, Eig_k. Then the $K \times K$ correlation matrix can also be written as,

$$Cor = \sum_{k=1}^{K} Lam_k(Eig_k)(Eig_k)'. \tag{9}$$

The larger eigenvalues explain more of the statistical dependence than the smaller ones, and hence another FOM is the eigenvalue-eigenvector pair,

$$\{Lam_1, Eig_1\} \tag{10}$$

followed by subsequent pairs $\{Lam_2, Eig_2\}$ and so forth.

In Section 2, we motivate the problem of prediction of x in a state-space model with a description of how atmospheric properties are retrieved from remote sensing data y. Section 3 presents FOMs based on the first two moments of the prediction error, including their use in simultaneous inference and how they can be obtained from a simulation experiment. Section 4 estimates the FOMs given by (6)–(8), in a simulation experiment involving a simple state-space model motivated by the exposition given in Section 2. Section 5 shows how simultaneous inference can be carried out on the multivariate state, and discussion and conclusions are given in Section 6.

2. Retrieving the State of the Atmosphere from Remote Sensing Data

Making inference on the hidden (or latent) state variables that generated a given dataset can be a challenging statistical problem, particularly when the forward model relating the data to the state is non-linear. Predicted values for the state variables are obtained by solving an "inverse problem"; the noise in the data makes the problem challenging. The solution depends on the form of the forward model, the assumed measurement-error (i.e., noise) model, and any prior information that is included in the analysis, all of which may vary over time or during calibration and testing of the model. Hence, there is a compelling need for easily interpretable FOMs, to determine the statistical properties associated with prediction of the state and to compare the effects of different modelling assumptions.

An important application of state-space modelling requiring the inverse problem to be solved, is the analysis of atmospheric remote sensing data. Here, a large number of sensor-based radiance measurements are used to infer relatively fewer underlying state variables. This inverse problem is ill-posed because the radiance measurements are noisy. Numerical approaches include those based on Twomey–Tikhonov regularisation (see Doicu et al., 2010, for a

recent review). Statistical approaches include ridge regression, penalised likelihood, and Bayesian posterior analysis (e.g., Cressie & Wang, 2013, and the references therein).

Column-averaged atmospheric carbon dioxide (CO_2) predictions are obtained for soundings from, for example, the Orbiting Carbon Observatory-2 (OCO-2) satellite and the Greenhouse gases Observing SATellite (GOSAT). NASA's retrieval algorithms take radiances y and solve the inverse problem using optimal estimation (Rodgers, 2000) in a Bayesian framework; a cost function is minimised to obtain predictions of the state x (e.g., Connor et al., 2008; O'Dell et al., 2012; Crisp et al., 2014). The prediction of the state is an iterative solution based on a Levenberg-Marquardt variant of a Gauss-Newton algorithm and enhanced with multiple filters and several processing steps. It is frequently tested (e.g., Bösch et al., 2011), updated (Eldering et al., 2014), and refined (e.g., Crisp et al., 2012; O'Dell et al., 2012) to improve the quality of its output. The result is a non-linear predictor, $\hat{x}(y)$, that cannot usually be written down in closed form.

Traditionally, *Bias* and *Mspe* given by (2) have been used to assess the predicted state variables obtained from a solution of the inverse problem, since they are the predominant FOMs. For the OCO-2 and GOSAT retrievals, estimates of *Bias* and *Mspe* are obtained from either a first-order (Connor et al., 2008) or a second-order (Cressie & Wang, 2013) Taylor-series approximation. However, given the complexity of the algorithm for obtaining the predictor $\hat{x}(y)$ of the state x from data y, simulation experiments are needed to obtain the actual (i.e., not the approximate) statistical properties of the prediction error (1).

3. Multivariate Inference: Use of Icv, Sdv, and Cor

From (2) and using iterated expectations, we obtain:

$$Bias = E_y\{E_{x|y}(\hat{x}(y)) - E_{x|y}(x)\} = E_y\{\hat{x}(y) - E_{x|y}(x)\} \tag{11}$$

$$\begin{aligned} Mspe &= E_y\{E_{x|y}(\hat{x}(y) - x)(\hat{x}(y) - x)'\} \\ &= E_y\{cov_{x|y}(\hat{x}(y) - x) + (E_{x|y}(\hat{x}(y)) - E_{x|y}(x))(E_{x|y}(\hat{x}(y)) - E_{x|y}(x))'\} \\ &= E_y\{cov_{x|y}(x)\} + E_y\{(\hat{x}(y) - E_{x|y}(x))(\hat{x}(y) - E_{x|y}(x))'\}. \end{aligned} \tag{12}$$

When $\hat{x}(y) = E_{x|y}(x)$, namely the posterior mean, we see from (11) and (12) that $Bias = 0$ and $Mspe = E_y\{cov_{x|y}(x)\}$. However, when other predictors such as those based on regularisation (see Section 2) are used, *Bias* is generally non-zero and *Mspe* should be calculated using (12).

3.1. Alternative Figures of Merit (FOMs)

Notice that whilst *Bias* is a function of the first moment of the prediction-error distribution, from (3), *Mspe* is a function of both the first moment and the second *central* moment. That is, prediction-error variability cannot be obtained only from *Mspe*, rather it requires *Cov* given by (3); see also (19) in Section 3.2 below. Moreover, for multivariate x, the relative magnitude of *Bias* and *Mspe* can vary by orders of magnitude for different state variables. Hence, the visualisation and interpretation of prediction regions may be influenced by the relative magnitude of the state variables. Consequently, from Section 1, alternative FOMs are:

$$Icv \equiv Sdv^{-1}Bias = Sdv^{-1}E_{x,y}(\hat{x}(y) - x) = Sdv^{-1}\{E_y(\hat{x}(y)) - E_x(x)\}, \tag{13}$$

$$Sdv \equiv (\mathrm{diag}(Cov))^{1/2} = \{\mathrm{diag}(\mathrm{cov}_{x,y}(\hat{x}(y) - x))\}^{1/2} = \{\mathrm{diag}[E_y(\mathrm{cov}_{x|y}(x)) + \mathrm{cov}_y(\hat{x}(y) - E_{x|y}(x))]\}^{1/2}, \tag{14}$$

and

$$Cor \equiv Sdv^{-1}(Cov)Sdv^{-1} = Sdv^{-1}[E_y(\mathrm{cov}_{x|y}(x)) + \mathrm{cov}_y(\hat{x}(y) - E_{x|y}(x))]Sdv^{-1}. \tag{15}$$

From (9) and (15), the symmetric square root of the inverse correlation matrix is,

$$Cor^{-1/2} \equiv \sum_{k=1}^{K} (Lam_k)^{-1/2} (Eig_k)(Eig_k)', \tag{16}$$

which is a $K \times K$ matrix we shall need in Section 3.2 below. Equivalently, $Cor^{-1/2} = (Eig)(Lam)^{-1/2}(Eig)$, which is well defined because $Lam_1 \geq Lam_2 \geq ... \geq Lam_k > 0$.

Notice that we can reconstruct the traditional FOMs from these new FOMs using,

$$Bias = (Sdv)(lcv) \tag{17}$$

$$Mspe = (Sdv)(Cor)(Sdv) + (Sdv)(lcv)(lcv)'(Sdv). \tag{18}$$

3.2. Simultaneous Inference

Recall that the prediction error is defined by (1). We make inference on the unknown state x through the prediction-error distribution,

$$(\hat{x}(y) - x) \sim \text{Dist}(\mu, \Sigma),$$

where "Dist" is a given distribution (e.g., the Gaussian distribution). The first two moments of "Dist," namely the prediction-error mean, μ, and the prediction-error covariance matrix, Σ, are given by

$$\mu = E_{y,x}(\hat{x}(y) - x) = Bias = (Sdv)(lcv)$$
$$\Sigma = \text{cov}_{y,x}(\hat{x}(y) - x) = Cov = Mspe - (Bias)(Bias)' = (Sdv)(Cor)(Sdv),$$

and "Dist" may depend on other parameters as well.

Notice that *lcv* is equal to μ scaled by Sdv^{-1}. Hence, it is a standardised or unit-free quantity that provides a direct bias comparison of prediction biases for all state variables, and it can be used to compare predictions from different factor combinations in a simulation experiment. As *Sdv* and *Cor* are matrix functions of the prediction-error covariance matrix, their elements measure the prediction variability and cross-dependence, respectively, and they may also be used to compare predictions from different factor combinations in a simulation experiment.

Assume "Dist" is *approximately* Gaussian, so that $\hat{x}(y) - x \dot{\sim} \text{Gau}(Bias, Cov)$, where recall that the multivariate state x is K-dimensional, and hence the Gaussian distribution is K-dimensional. Let $c \equiv \chi^2_K(0.95)$ denote the upper 95-th percentile of a chi-squared distribution on K degrees of freedom. Then the Wald statistic,

$$W^2 \equiv (\hat{x}(y) - x - Bias)' Cov^{-1} (\hat{x}(y) - x - Bias),$$

is approximately χ^2_K-distributed, and so, approximately,

$$Pr(W^2 \leq c) = 0.95.$$

Since $Cov = Sdv(Cor)Sdv$, an approximate 95% *prediction ellipsoid* for the state x is derived as follows:

$$\begin{aligned} Ell(0.95) &\equiv \{x : (x - \hat{x}(y) + Bias)' Cov^{-1} (x - \hat{x}(y) + Bias) \leq c\} \\ &= \{x : (x - \hat{x}(y) + Bias)' Sdv^{-1} Cor^{-1/2} Cor^{-1/2} Sdv^{-1} (x - \hat{x}(y) + Bias) \leq c\} \\ &= \{x : (\omega - \hat{\omega}(y) + Cor^{-1/2} lcv)'(\omega - \hat{\omega}(y) + Cor^{-1/2} lcv) \leq c\}, \end{aligned} \tag{19}$$

where

$$\omega \equiv Cor^{-1/2} Sdv^{-1} x; \qquad \hat{\omega}(y) \equiv Cor^{-1/2} Sdv^{-1} \hat{x}(y). \tag{20}$$

Hence, in terms of the transformed variables ω and $\hat{\omega}(y)$,

$$Sph(0.95) \equiv \{\omega : (\omega - \hat{\omega}(y) + Cor^{-1/2} lcv)'(\omega - \hat{\omega}(y) + Cor^{-1/2} lcv) \leq c\}, \tag{21}$$

is an approximate 95% *prediction spheroid* centred at $(\hat{\omega}(y) - Cor^{-1/2} lcv)$ with radius $c = \chi^2_K(0.95)$ such that, approximately, $\Pr(Sph(0.95)) = 0.95$.

Note the absence of units in (20) and (21). The vectors ω in the sphere define a 95% prediction region in the state space through $x = (Sdv)(Cor^{1/2})\omega$, where the units of x are recovered after ω is rescaled. Thus the FOMs *lcv*, *Sdv*, and *Cor* have an easily interpretable role in simultaneous inference on the state x; and simultaneous inference given by (19), which can be derived from (21), is well known to be more efficient than inferring individual state elements one-at-a-time; see Section 5 for an illustration of this.

3.3. Obtaining FOMs from a Simulation Experiment

Ideally, FOMs can be obtained analytically however, for nonlinear statistical models, closed-form expressions are rarely available. Consider instead a simulation experiment where a state-space model is used to generate L independent replications of

State:	$x^{(1)}, \ldots, x^{(L)}$
Data:	$y^{(1)}, \ldots, y^{(L)}$
Predictor:	$\hat{x}(y^{(1)}), \ldots, \hat{x}(y^{(L)})$
Prediction error:	$\hat{x}(y^{(1)}) - x^{(1)}, \ldots, \hat{x}(y^{(L)}) - x^{(L)}$.

The statistical properties of the prediction error and various FOMs can be *estimated* from the simulation experiment. For example, the *Bias* and *Mspe* referred to in (2) are estimated unbiasedly by,

$$\hat{Bias} \equiv \frac{1}{L} \sum_{l=1}^{L} (\hat{x}(y^{(l)}) - x^{(l)}) \; ; \qquad \hat{Mspe} \equiv \frac{1}{L} \sum_{l=1}^{L} (\hat{x}(y^{(l)}) - x^{(l)})(\hat{x}(y^{(l)}) - x^{(l)})'. \tag{22}$$

As L increases, the estimate, $\hat{Cov} \equiv \hat{Mspe} - (\hat{Bias})(\hat{Bias})'$, is an asymptotically unbiased estimate of *Cov*. Likewise,

$$\hat{lcv} \equiv \hat{Sdv}^{-1}(\hat{Bias}) \; ; \qquad \hat{Sdv} \equiv (\mathrm{diag}(\hat{Cov}))^{1/2}; \qquad \hat{Cor} \equiv \hat{Sdv}^{-1}(\hat{Cov})\hat{Sdv}^{-1}, \tag{23}$$

are asymptotically unbiased estimates of the FOMs *lcv*, *Sdv*, and *Cor*, respectively.

4. Simulations from a Bivariate State-Space Model

In this section, we illustrate how our proposed FOMs can be used for evaluating a simulation experiment. Our motivation for this simple experiment is prediction of the state of the atmosphere based on remote sensing data; here we use simulated radiance measurements to predict the volume mixing ratios of CO_2 and of O_2. However, we have simplified the problem greatly by assuming that the radiances were obtained for a notional vertical column of the atmosphere between 4.5 km and 5.5 km in altitude, free of aerosols and hence with a pathlength of $ds = 2$ km. In this partial

column of the atmosphere, pressure decreases approximately linearly with height, and hence a height of $s_m = 5.5$ km, with temperature $T_m = 252.43$ K and pressure $P_m = 0.4988$ atm was considered representative of the partial column. Analogous conceptualisation and simulations can be found in the marine-sciences literature, where processes interact in a notional "mixing zone"; there, the models upon which such simulations are based are referred to as "box models" (e.g., Parslow et al., 2013). Further, in our experiment we assumed that solar flux, I_0, and reflectance, R, had constant, given values; that nadir data were obtained; and that changes in radiance were only due to absorption (i.e., we ignored emission and scattering effects).

We varied two factors in the simulation experiment: the correlation ρ between the two states of the atmosphere and the signal-to-noise ratio, $SNR \equiv \sqrt{\sum_{j=1}^{20} \text{var}(F_j(x))}/\sqrt{\text{tr}(\Sigma_\epsilon)}$, where $\text{var}(F_j(x))$ is the empirical variance of the forward function $F_j(x)$, for $j = 1, ..., 20$ (defined below), calculated using simulated realisations of the bivariate state x; and Σ_ϵ is the measurement-error covariance matrix (defined below). Two levels were specified for each factor (i.e., two values of ρ and two values of SNR), and hence there were $2 \times 2 = 4$ combinations to compare.

The fundamental element of our experiment is a vector-valued simulation of 20 radiances, y, obtained using a forward function defined in terms of the bivariate state vector, x. The state vector consists of the mole fractions of CO_2 and O_2, and it was generated using a bivariate Gaussian distribution. From this x, the 20-dimensional data vector y was generated using a forward function (defined below) and an additive, Gaussian, measurement-error term. We then used optimal estimation (defined below) to solve the inverse problem, and hence we obtained a prediction $\hat{x}(y)$ for x. This simulation was repeated many times for each factor combination, resulting in Monte-Carlo, method-of-moment estimates for *Bias*, *Mspe*, *Icv*, *Sdv*, and *Cor* of the prediction-error distribution; see (22) and (23).

4.1. The State-Space Model

In our simple simulation experiment, two state variables, x_1 and x_2, represent the volume mixing ratios of CO_2 and of O_2, respectively, in units of parts per million (ppm). In general, the state $x = (x_1, x_2)'$ cannot be measured directly, so here we inferred it from simulated remote sensing radiance measurements $I(\cdot)$, calculated using a simplified radiative transfer equation, for a set of wavenumbers $\{\nu_i : i = 1, ..., n\}$ chosen from three regions of the spectrum: The strong CO_2 band (4810 - 4897 cm^{-1}), the weak CO_2 band (6170 - 6270 cm^{-1}), and the O_2 A-band (12950 - 13190 cm^{-1}).

To calculate the radiance at wavenumber ν_i, we obtained (from the HITRAN2012 database; see Rothman et al., 2013) the pressure-shift constant $\delta_{km}(\nu_i)$, the lower-state energy $E_k''(\nu_i)$, the line strength $s_k(\nu_i, T_0)$ at reference temperature T_0, the air-broadened half width $\alpha_k(\nu_i, T_0, P_0)$, the temperature exponent γ for a reference temperature $T_0 = 296$ K and pressure $P_0 = 1013.2$ hectopascals $= 1.0$ atm, and the zero-pressure line centre ν_{i0}^0. Then, we used the line centre at pressure P_m, namely $\nu_{im}^0 = \nu_{i0}^0 + (P_m/P_0)\delta_{km}(\nu_i)$; the second radiation constant,$C_2 = hc/K_B$, where h is Planck's constant, K_B is Boltzmann's constant, and c is the speed of light; and the total internal partition sum (TIPS), $Q_{k,t}(T_m)$, from the TIPS module of the HITRAN2012 database, to calculate the line strength at temperature T_m, for each wavenumber ν_i, given by,

$$s_k(\nu_i, T_m) = s_k(\nu_i, T_0)\frac{Q_{k,t}(T_0)(1 - \exp(-C_2\nu_{im}^0/T_m))}{Q_{k,t}(T_m)(1 - \exp(-C_2\nu_{im}^0/T_0))} \exp\left(-C_2 E_k''(\nu_i)\left(\frac{1}{T_m} - \frac{1}{T_0}\right)\right).$$

Ignoring the self-broadened half-width, the air-broadened half-width at temperature T_m and pressure P_m, for each wavenumber (ν_i), is given by,

$$\alpha_k(\nu_i, T_m, P_m) \simeq \alpha_k(\nu_i, T_0, P_0)\frac{P_m}{P_0}\left(\frac{T_0}{T_m}\right)^\gamma,$$

and the optical mass of state variable x_k, is given by,

$$N(x_k, T_m, P_m) = \frac{x_k P_m ds}{K_B T_m} = x_k N(T_m, P_m).$$

Since all altitudes are less than 16 km, the Lorentz-line-shape function is appropriate. Hence, the absorption cross-section of state x_k, at wavenumber ν_i, temperature T_m, and pressure P_m, is given by,

$$K_k(\nu_i, T_m, P_m) = \frac{s_k(\nu_i, T_m)}{\pi} \frac{\alpha_k(\nu_i, T_m, P_m)}{(\nu_i - \nu_i^0)^2 + \alpha_k(\nu_i, T_m, P_m)^2}.$$

Finally, using the Beer-Lambert law, the total radiance for each wavenumber, ν_i, is given by,

$$I(\nu_i) = I_0(\nu_i) R(\nu_i, \theta) \exp\left(- \sum_{k=1}^{2} K_k(\nu_i, T_m, P_m) N(x_k, T_m, P_m) \right), \tag{24}$$

where for simplicity, T_m and P_m are dropped from the notation on the left hand side. In our simulation experiment, the solar flux $I_0(\nu_i)$, the reflectance $R(\nu_i, \theta)$, and the angular parameters θ, had constant values.

For the three spectral bands used in this experiment, CO_2 and O_2 parameter values are available from the HITRAN2012 database for a total of 17279 CO_2 wavelengths and 446 O_2 wavelengths. In the CO_2 weak and strong bands, the OCO-2 spectrometer has a resolution of approximately 0.262 cm^{-1} and 0.258 cm^{-1}, respectively. This motivated us to divide the CO_2 strong and weak bands into 333 and 466 intervals, respectively, and we averaged the radiances in each interval using a weighted average, with weights defined by the relative abundance of each CO_2 and O_2 isotopologue. We used the 446 individual wavelengths in the O_2 A-band. Hence, for a spectral interval $d\nu_j$ centred at ν_j that includes n_j wavenumbers $\{\nu_{jr} : r = 1, ..., n_j\}$, the weighted-average radiance is y_j and the nonlinear forward function $F_j(x)$ is given by,

$$F_j(x) = C_{IR} \exp\left(- \sum_{k=1}^{2} x_k \sum_{r=1}^{n_j} w(\nu_{jr}) K_k(\nu_{jr}, T_m, P_m) N(P_m, T_m) \right); \quad x = (x_1, x_2)', \tag{25}$$

where C_{IR} is a constant that approximates the solar flux and reflectance parameters, and $w(\nu_{jr})$ is the normalised relative abundance of the isotopologue at wavenumber ν_{jr}. For this simulation experiment, there were $333 + 466 + 446 = 1255$ wavenumber intervals under consideration (Figure 1). From these, we selected $J = 20$ wavenumber intervals to use in the experiment: there were seven each from the CO_2 strong and weak bands and six from the O_2 A-band.

To simulate radiances using (25), we first defined a distribution for $x = (x_1, x_2)'$:

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \sim \text{Gau}\left(\mu_x = \begin{pmatrix} 390 \\ 209550 \end{pmatrix}, \Sigma_x = \begin{bmatrix} 4 & \sigma_{12} \\ \sigma_{12} & 16 \end{bmatrix} \right), \tag{26}$$

where $\sigma_{12} = (4 \times 16)^{1/2} \rho = 8\rho$. The correlation coefficient ρ was a factor in the experiment, with two levels: $\rho = -0.2$ and $\rho = -0.8$. Using (26), we simulated $L = 20,000$ realisations of the state variables, $x^{(l)} = (x_1^{(l)}, x_2^{(l)})'$, $l = 1, ..., 20,000$, and for each l we simulated a data vector of $J = 20$ radiances $y^{(l)} = (y_1^{(l)}, ..., y_{20}^{(l)})'$ using,

$$y_j^{(l)} = F_j(x^{(l)}) + \epsilon_j^{(l)}, \tag{27}$$

where realisations of the measurement error, $\epsilon_j^{(l)}$, were independently distributed as $\epsilon_j^{(l)} \sim \text{Gau}(0, \sigma_\epsilon^2)$, for $j = 1, ..., 20$. Then the measurement-error covariance is the 20×20 diagonal matrix $\Sigma_\epsilon = \sigma_\epsilon^2 I$, where I is the 20×20 identity matrix. We selected two levels for σ_ϵ^2 corresponding to $SNR = 0.5$ and $SNR = 2.0$. These signal-to-noise ratios are lower than those obtained from, say, retrievals of a 48-dimensional state vector based on data from the OCO-2 satellite; they were chosen in order to compensate for the simplicity of the bivariate state-space model used in the experiment.

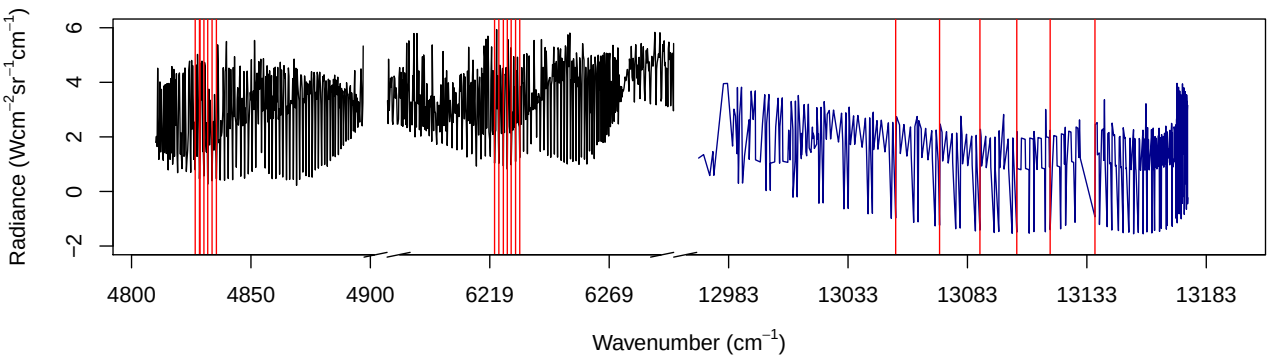


Figure 1. One realisation of the 1255 simulated radiance measurements for the strong CO_2 and weak CO_2 bands, and the O_2 A-band. A red line indicates the $7+7+6=20$ wavenumber intervals selected from the three bands for the simulation experiment. The 20 radiances $(y_1, ..., y_{20})$ represent the data used from a single sounding to infer the underlying true state (x_1, x_2). These are the data used for simultaneous inference in Section 5.

For each $y^{(l)}$, $l = 1, ..., 20,000$, we used optimal estimation (Rodgers, 2000) to obtain the predictor, $\hat{x}(y^{(l)})$, by minimising the cost function,

$$(y^{(l)} - F(x))'\Sigma_\epsilon^{-1}(y^{(l)} - F(x)) + (x - \mu_x)'\Sigma_x^{-1}(x - \mu_x), \tag{28}$$

with respect to x. In our case, the number of states was only two, and hence we were able to use a general-purpose optimisation algorithm from the R statistical software. The optimisation routine we implemented used a limited-memory modification of the quasi-Newton method and a finite-difference approximation for the gradient. To avoid undefined solutions, we set the lower bound for each x_1 and x_2 to a very small number. Minimising (28) for each $l = 1, ..., 20,000$, resulted in the 20,000 independent predictors of the state, $\{(\hat{x}_1^{(l)}, \hat{x}_2^{(l)}) : l = 1, ..., 20,000\}$.

4.2. Results

The prediction error from the l-th simulation, $\hat{x}(y^{(l)}) - x^{(l)}$, for $l = 1, ..., 20,000$, is two-dimensional, where recall that $x^{(l)}$ is the (simulated) true value of the state for the l-th simulation. Histograms and a density plot (Figure 2) of the prediction errors illustrate the variability in the prediction errors for CO_2 and O_2 for one combination of factor levels. For each of the four factor-level combinations, we obtained the traditional FOMs, $\hat{B}ias$ and $\hat{M}spe$, according to (22) and the FOMs, $\hat{l}cv$, $\hat{S}dv$, and $\hat{C}or$ according to (23). From $\hat{C}or$ we also obtained $\hat{L}am_k$ and $\hat{E}ig_k$, for $k = 1, 2$.

Figure 3 compares $\hat{B}ias$ with $\hat{l}cv$, for CO_2 and O_2 and for each factor-level combination in our simulation experiment. The change in relative magnitudes from $\hat{B}ias$ to $\hat{l}cv$ is particularly striking for prediction of CO_2. For O_2, the greatest $\hat{l}cv$ occurs when $SNR = 0.5$ and $\rho = -0.8$; for CO_2 it occurs when $SNR = 2.0$ and $\rho = -0.2$.

Figure 4 shows that the magnitudes of both $\hat{M}spe$ and $\hat{S}dv$ are greater for O_2 than for CO_2. For O_2, $SNR = 2.0$ and $\rho = -0.8$ gives the smallest $\hat{M}spe$ and $\hat{S}dv$ values.

The correlation between the prediction errors for CO_2 and O_2 (Table 1) is low when $\rho = -0.2$, for both levels of SNR. When $\rho = -0.8$, the prediction-error correlation is more substantial, particularly for $SNR = 0.5$. For a bivariate correlation matrix, $\hat{E}ig_1$ and $\hat{E}ig_2$ are the same for each factor-level combination; hence we consider only the eigenvalues, $\hat{L}am_k$, $k = 1, 2$. For a given factor-level combination, the eigenvalues $\hat{L}am_1$ and $\hat{L}am_2$ (Figure 5) have magnitudes

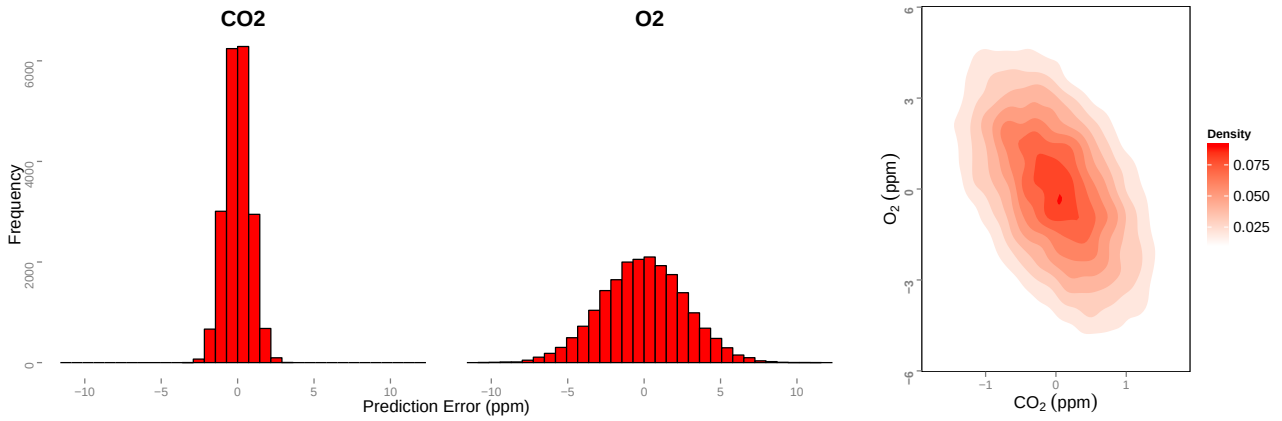


Figure 2. Univariate and bivariate density estimates of the prediction-error distribution for CO_2 and O_2, obtained from the simulation experiment described in Section 4 with $SNR = 0.5$ and $\rho = -0.8$. The counts in the univariate density estimates are out of 20,000, and all other axes are in units of ppm.

Table 1. Correlation between the prediction errors for CO_2 and O_2, given by the element $\hat{C}or_{1,2}$ from the 2×2 matrix $\hat{C}or$, for each factor-level combination of the simulation experiment described in Section 4.

	$SNR = 0.5$	$SNR = 2.0$
ρ=-0.2	-0.081	-0.021
ρ=-0.8	-0.478	-0.158

close to 1 when there is low prediction-error correlation. Greater difference between $\hat{L}am_1$ and $\hat{L}am_2$ are associated with more substantial prediction-error correlation.

5. Simultaneous Inference on the State Elements

In this section, we illustrate the use of FOMs for simultaneous inference. Data are required to make inference on the state variables, which would be obtained from atmospheric remote sensing measurements. In our case, we obtained a data vector y by first simulating one new realisation of the hidden 'true' state; using (26) and $\rho = -0.8$, we obtained the simulated true state $x = (392.43, 209545.9)'$. Then, using the forward model (27) with $SNR = 0.5$, we generated a vector of radiances y, whose elements are indexed by the wavelengths at the vertical lines shown in Figure 1. This is the data vector to which the methodology outlined in Section 4 was applied to obtain a prediction $\hat{x}(y) = (391.37, 209547.8)'$. In general, the true state is unknown; in what follows, we make inference on x, using two prediction regions obtained from univariate prediction intervals, and using a simultaneous prediction region based on (19), derived from (21). The latter is found to have superior statistical properties.

We first consider inference on the state variables separately. A 95% univariate prediction interval for each state variable (e.g., x_1) is given by,

$$Pr\left(\frac{|\hat{x}_1 - x_1 - Bias_1|}{(Cov_{11})^{1/2}} < c_u\right) = 0.95, \tag{29}$$

where c_u is the upper 97.5-th percentile of a univariate Gaussian distribution. Approximately 95% of the time, the true state will lie inside the interval defined in (29); see Figures 6a and 6b. As the 'true' state is known in this case, we

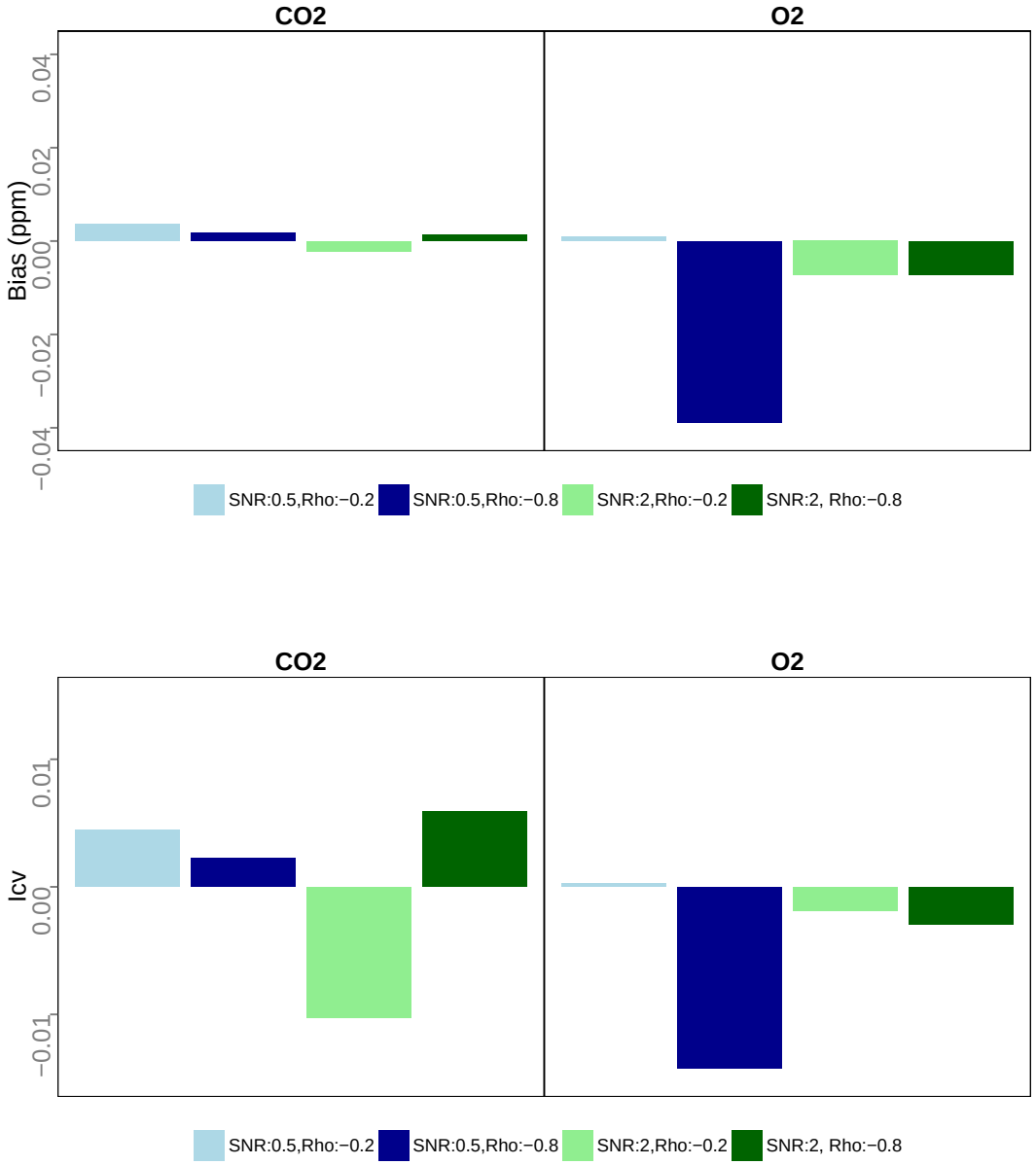


Figure 3. Bar charts of $\hat{Bias}$ given by (22) (top panel), and of $\hat{lcv}$ given by (23) (bottom panel), of the prediction-error distribution for CO_2 and O_2, obtained for each factor-level combination from the simulation experiment described in Section 4. Notice the different vertical scales in the two panels.

can observe its location in the prediction interval. Notice that for our simulated realisation, the univariate prediction

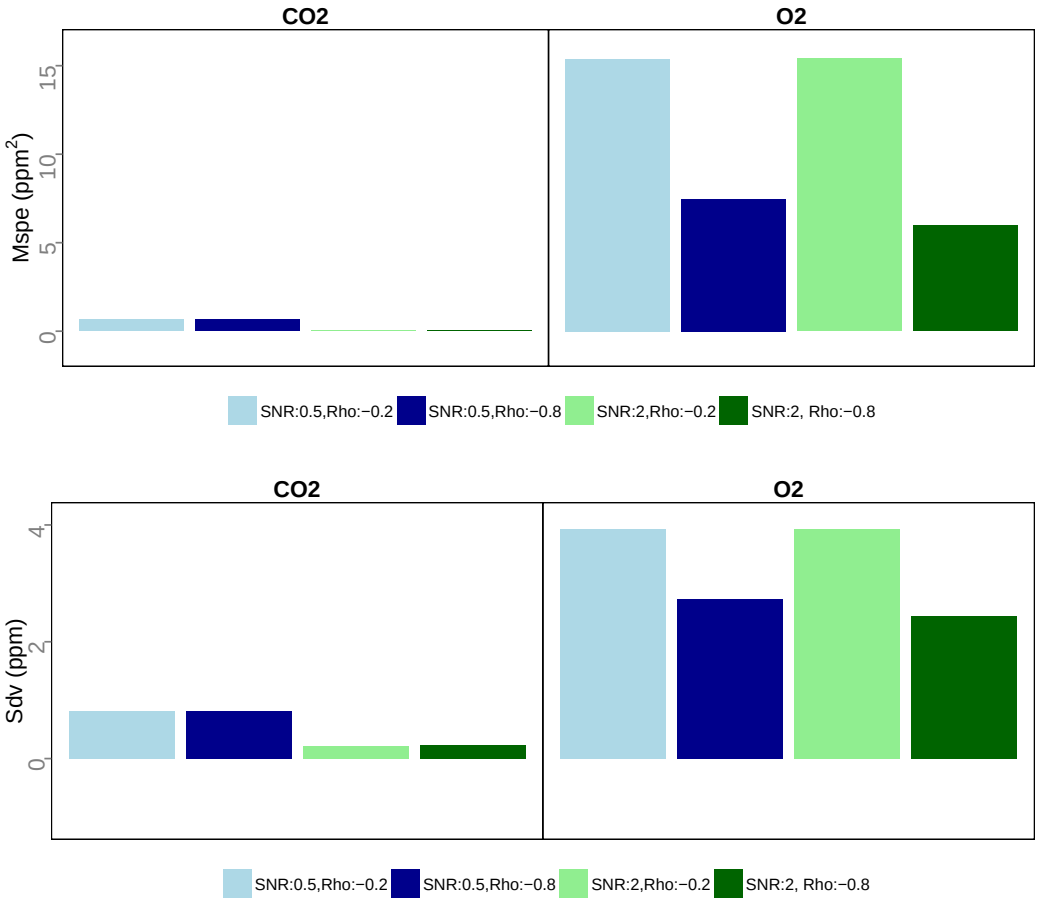


Figure 4. Bar charts of $\hat{M}spe$ given by (22) (top panel), and of $\hat{S}dv$ given by (23) (bottom panel), of the prediction-error distribution for CO_2 and O_2, obtained for each factor-level combination from the simulation experiment described in Section 4. Notice the different vertical scales in the two panels.

intervals are $(389.77, 392.97)$ for CO_2, and $(209542.5, 209553.2)$ for O_2. These intervals ignore the presence of other state variables, so whilst each interval contains 95% probability, from the simulation in Section 4 we obtain $Pr\left(\left\{\frac{|\hat{x}_1 - x_1 - Bias_1|}{(Cov_{11})^{1/2}} < c_u\right\} \cap \left\{\frac{|\hat{x}_2 - x_2 - Bias_2|}{(Cov_{22})^{1/2}} < c_u\right\}\right) = 0.908$; see Figure 6c.

Alternatively, we can use simultaneous prediction intervals that are adjusted according to the dimensionality of the state vector (e.g., using a Bonferroni adjustment; see Figure 6d). Here, a Bonferroni-adjusted nominal 95% simultaneous prediction interval for x satisfies,

$$Pr\left(\left\{\frac{|\hat{x}_1 - x_1 - Bias_1|}{(Cov_{11})^{1/2}} < c_b\right\} \cap \left\{\frac{|\hat{x}_2 - x_2 - Bias_2|}{(Cov_{22})^{1/2}} < c_b\right\}\right) \geq 0.95, \tag{30}$$

where c_b is the Bonferroni-adjusted critical value for a univariate Gaussian distribution. In our case, the number of state variables is 2, and hence $1 - 0.95 = 0.05$ is divided by 2, resulting in 0.025. We divide by 2 again to account for the symmetry of the Gaussian distribution, and hence c_b is the upper $1 - 0.0125 = 98.75$-th percentile of a

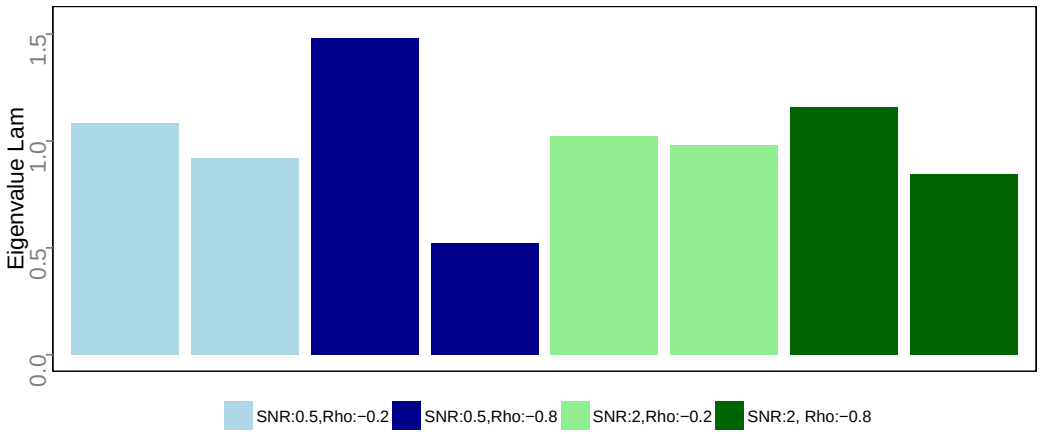


Figure 5. The first two eigenvalues, $\hat{L}am_1$ and $\hat{L}am_2$, of $\hat{C}or$, given by (23), for each factor-level combination from the simulation experiment described in Section 4.

univariate Gaussian distribution. Notice that the prediction intervals in Figure 6d, of $(389.54, 393.20)$ for CO_2 and $(209541.7, 209554.0)$ for O_2, are wider than the individual univariate prediction intervals in Figures 6a and 6b. The joint probability of being in the intervals is $Pr(\{\frac{|\hat{x}_1 - x_1 - Bias_1|}{(Cov_{11})^{1/2}} < c_b\} \cap \{\frac{|\hat{x}_2 - x_2 - Bias_2|}{(Cov_{22})^{1/2}} < c_b\}) = 0.953$, which as expected is larger than 0.95.

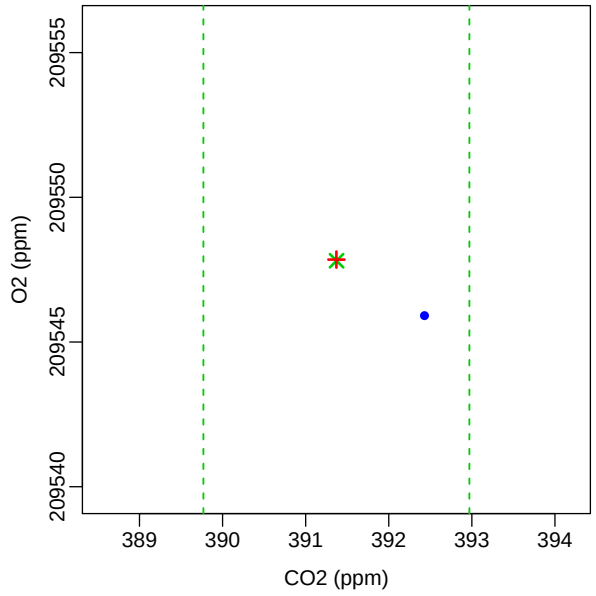
Now consider simultaneous inference for the multivariate state. Using a 95% prediction spheroid based on *lcv*, *Sdv* and *Cor*, given by (21), we obtain Figure 7a. The idea is that the prediction spheroid is easy to construct and, from (20), a simple back-transformation results in the corresponding 95% prediction ellipsoid given by (19) and shown in Figure 7b.

This section demonstrates the efficiency of simultaneous inference versus inference based on univariate prediction intervals. The prediction region in Figure 7 has the joint distribution of the multivariate state as its basis. Hence, when the respective prediction errors are not independent, the prediction region for CO_2 (O_2) is narrower for a given value of O_2 (CO_2) than the corresponding univariate prediction interval for CO_2 (O_2), thus improving the efficiency of inference on the multivariate state; see Figures 7c and 7d.

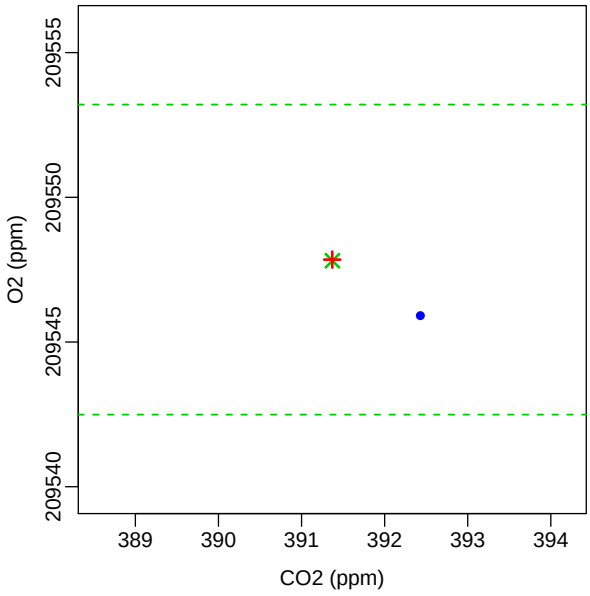
6. Discussion and Conclusions

FOMs obtained from statistical properties of the multivariate estimation error or the multivariate prediction error are important both for evaluating simulation experiments and for simultaneous inference. *Bias* and *Mspe* are traditionally used and, whilst their use as FOMs is widely accepted, they have some limitations. In this article, we propose *lcv*, *Sdv*, and *Cor* as alternatives to *Bias* and *Mspe* for visualisation, evaluation, and interpretation of simulation experiments. We further show their role in inference on the unknown state variables, particularly the construction of simultaneous prediction regions. The exposition in this article is given for inference on the whole state vector x, but everything carries over to inference on a subvector of smaller dimension. Then the appropriate FOMs are defined in terms of the marginal multivariate distribution of the subvector of prediction errors.

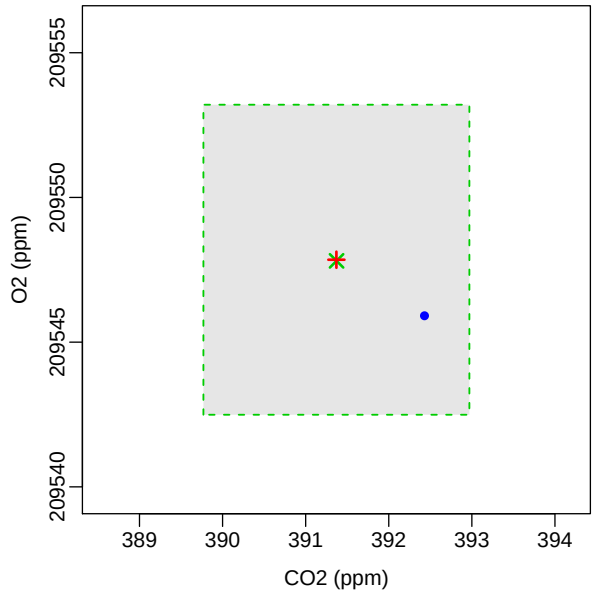
There are situations where we require statistical properties of the prediction error, $\hat{x}(y) - x$, conditional on the data y. In Section 4, this would correspond to inferring CO_2 and O_2 at a single sounding that resulted in obtaining the data



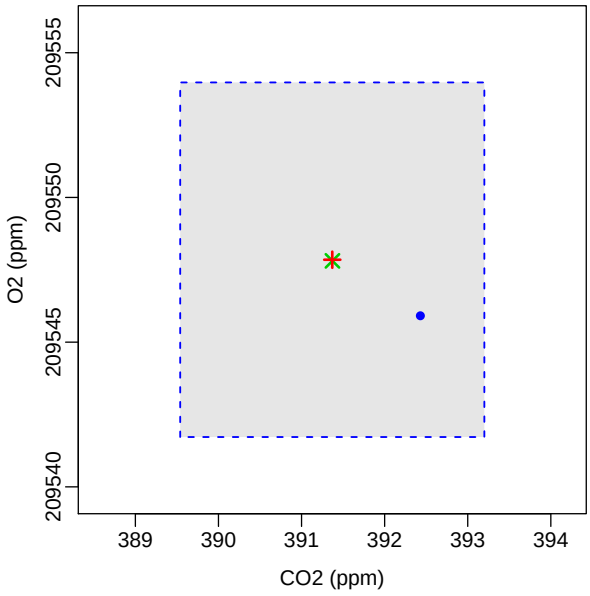
(a) Univariate 95% prediction interval for CO_2 given by (29).



(b) Univariate 95% prediction interval for O_2 given by (29).

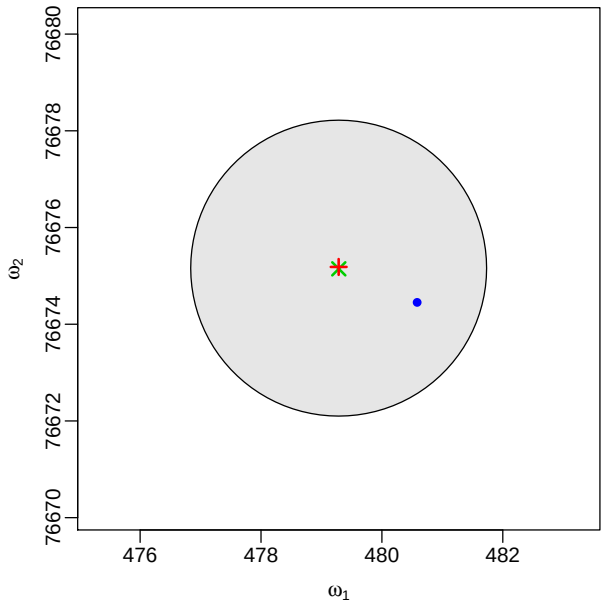


(c) Nominal 95% prediction region for CO_2 and O_2 obtained from univariate prediction intervals for CO_2 and O_2 given by (29).

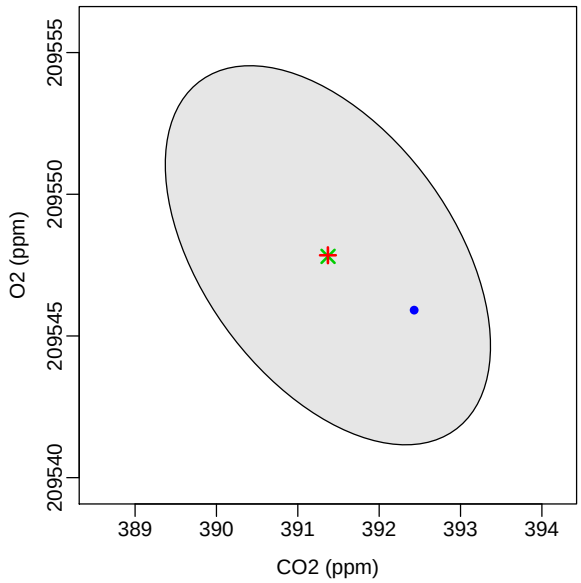


(d) Bonferroni-adjusted nominal 95% simultaneous prediction region given by (30).

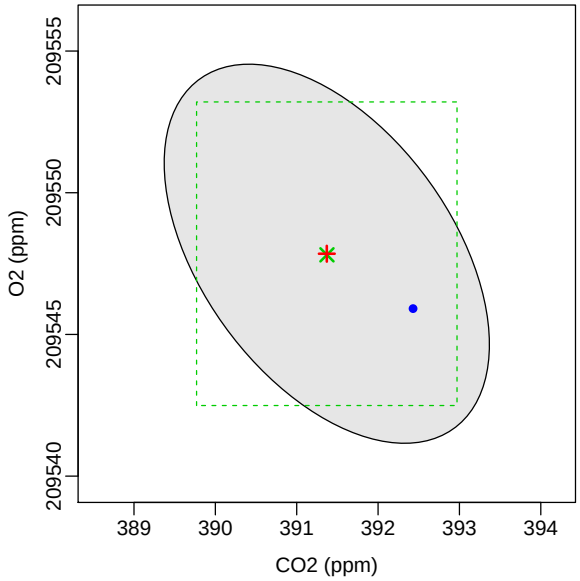
Figure 6. Prediction intervals and prediction regions for state variables CO_2 and O_2, obtained from a single (simulated) data vector y. The state-space model from which the data vector is simulated is described in Section 4 and has SNR=*0.5* and ρ=$-$*0.8*. Key: × prediction; + bias; ----- Region from univariate prediction intervals; ----- Bonferroni-adjusted prediction region; • true state.



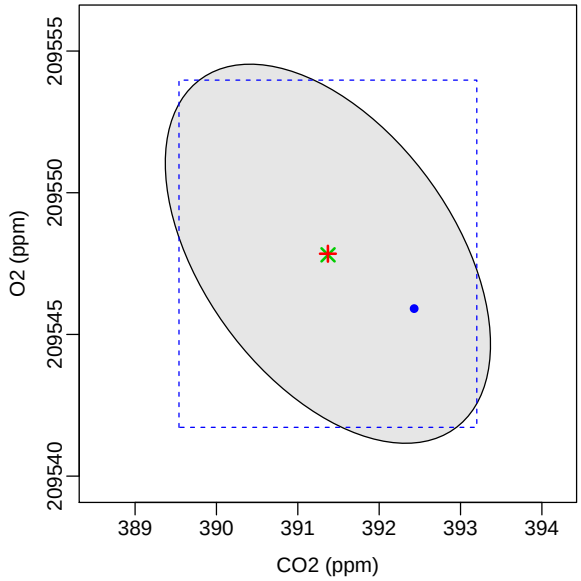
(a) The 95% prediction spheroid for transformed CO_2 and O_2, given by (21).



(b) The corresponding 95% prediction ellipsoid for CO_2 and O_2, given by (19).



(c) The 95% prediction ellipsoid for CO_2 and O_2, given by (19), and a nominal 95% simultaneous prediction region obtained from univariate prediction intervals for CO_2 and O_2, given by (29).



(d) The 95% prediction ellipsoid, given by (19), and a Bonferroni-adjusted nominal 95% simultaneous prediction region for CO_2 and O_2, given by (30).

Figure 7. The 95% prediction spheroid given by (21) and the corresponding 95% prediction ellipsoid given by (19) for multivariate prediction of state variables CO_2 and O_2, obtained from a single (simulated) data vector y. The state-space model from which the data vector is simulated is described in Section 4 and has SNR=*0.5* and ρ=*−0.8*. Key: × prediction; + bias; ------ Region from univariate prediction intervals; ------ Bonferroni-adjusted prediction region; —— simultaneous prediction regions; • true state.

vector y. These properties are immediately available from the predictive distribution, $p(x|y) \propto f(y|x)g(x)$, which can be found using *inter alia* Markov chain Monte Carlo.

Acknowledgements

This research was partially supported by the National Aeronautics and Space Administration (NASA) grant NNH11ZDA001N-OCO2 (OCO-2 Science Team) and NASA grant NNH11ZDA001N-AIST (Advanced Information Systems Technology); and it was partially supported by a 2015-2017 Australian Research Council Discovery Grant. We would like to acknowledge early discussions with Mike Gunson, Jon Hobbs, and Amy Braverman on a surrogate forward model that inspired the simple state-space model given in Section 4 and the role of the FOMs in simulation experiments. Our thanks also go to Nicholas Deutscher for his advice on the forward-model parameters used in Section 4.

References

Bösch, H, Connor, B, Crisp, D & Miller, C (2011), 'Global characterization of CO2 column retrievals from shortwave-infrared satellite observations of the Orbiting Carbon Observatory-2 Mission,' *Remote Sensing*, **3**, pp. 270–304.

Connor, BJ, Bösch, H, Toon, G, Sen, B, Miller, C & Crisp, D (2008), 'Orbiting Carbon Observatory: Inverse method and prospective error analysis,' *Journal of Geophysical Research: Atmospheres*, **113**, pp. 1–14.

Cressie, N & Wang, R (2013), 'Statistical properties of the state obtained by solving a nonlinear multivariate inverse problem,' *Applied Stochastic Models in Business and Industry*, **29**, pp. 424–438.

Crisp, D, Bösch, H, Brown, L, Castano, R, Christi, M, Connor, B, Eldering, A, Fisher, B, Frankenberg, C, Gunson, M, McDuffie, J, Miller, CE, Natraj, V, O'Dell, C, O'Brien, D, Polonski, I, Osterman, GB, Oyafuso, F, Smyth, M, Thompson, D, Toon, G & Spurr, R (2014), 'OCO (Orbiting Carbon Observatory)-2 Level 2 Full Physics Retrieval Algorithm Theoretical Basis,' Pasadena, CA.

Crisp, D, Fisher, BM, O'Dell, C, Frankenberg, C, Basilio, R, Bösch, H, Brown, LR, Castano, R, Connor, B, Deutscher, NM, Eldering, A, Griffith, D, Gunson, M, Kuze, A, Mandrake, L, McDuffie, J, Messerschmidt, J, Miller, CE, Morino, I, Natraj, V, Notholt, J, O'Brien, DM, Oyafuso, F, Polonsky, I, Robinson, J, Salawitch, R, Sherlock, V, Smyth, M, Suto, H, Taylor, TE, Thompson, DR, Wennberg, PO, Wunch, D & Yung, YL (2012), 'The ACOS CO_2 retrieval algorithm – Part II: Global X_{CO_2} data characterization,' *Atmospheric Measurement Techniques*, **5**, pp. 687–707.

Doicu, A, Trautmann, T & Schreier, F (2010), *Numerical Regularization for Atmospheric Inverse Problems*, Springer, Berlin and London.

Eldering, A, Pollock, R, Lee, RAM & Rosenberg, R (2014), 'Orbiting Carbon Observatory (OCO)-2 Level 1B Theoretical Basis Document,' Jet Propulsion Laboratory, Pasadena, CA.

O'Dell, CW, Connor, B, Bösch, H, O'Brien, D, Frankenberg, C, Castano, R, Christi, M, Eldering, D, Fisher, B, Gunson, M, McDuffie, J, Miller, CE, Natraj, V, Oyafuso, F, Polonsky, I, Smyth, M, Taylor, T, Toon, GC, Wennberg, PO & Wunch, D (2012), 'The ACOS CO_2 retrieval algorithm – Part 1: Description and validation against synthetic observations,' *Atmospheric Measurement Techniques*, **5**, pp. 99–121.

Parslow, J, Cressie, N, Campbell, EP, Jones, E & Murray, L (2013), 'Bayesian learning and predictability in a stochastic nonlinear dynamical model,' *Ecological Applications*, **23**, pp. 679–698.

Rodgers, CD (2000), *Inverse Methods for Atmospheric Sounding: Theory and Practice*, World Scientific, Singapore.

Rothman, LS, Gordon, IE, Babikov, Y, Barbe, A, Chris Benner, D, Bernath, PF, Birk, M, Bizzocchi, L, Boudon, V, Brown, LR, Campargue, A, Chance, K, Cohen, EA, Coudert, LH, Devi, VM, Drouin, BJ, Fayt, A, Flaud, JM, Gamache, RR, Harrison, JJ, Hartmann, JM, Hill, C, Hodges, JT, Jacquemart, D, Jolly, A, Lamouroux, J, Le Roy, RJ, Li, G, Long, DA, Lyulin, OM, Mackie, CJ, Massie, ST, Mikhailenko, S, Müller, H, Naumenko, OV, Nikitin, AV, Orphal, J, Perevalov, V, Perrin, A, Polovtseva, ER, Richard, C, Smith, M, Starikova, E, Sung, K, Tashkun, S, Tennyson, J, Toon, GC, Tyuterev, V & Wagner, G (2013), 'The HITRAN2012 molecular spectroscopic database,' *Journal of Quantitative Spectroscopy and Radiative Transfer*, **130**, pp. 4–50.

Santner, TJ, Williams, BJ & Notz, WI (2003), *The Design and Analysis of Computer Experiments*, Springer, New York, NY.