Post-processing multimodel ensembles: A multilevel factor analysis approach

Stefan Siegert

August 3, 2016

Contents

1 Introduction

multiple ensembles for the same prediction target; ensemble of ensembles; superensemble

MMEs are highly structured data with a (potentially) very rich correlation structure (fill in some details and examples)

Questions: How to compare performance of models within an MME? How to optimally combine the MME into a single prediction?

We propose a statistical framework to model within-model and between-model variability (variance decomposition)

1.1 Literature

multilevel factor analysis http://dx.doi.org/10.1207/s15327752jpa8402_02

two-level framework (model, member)

Krishnamurty 1999: Multiple regression for MME http://

science.sciencemag.org/content/285/5433/1548.full.pdf+html

Afshartous 2005: Prediction in multi-level models http://jeb.sagepub.com/content/30/2/109.short "The multilevel prediction rule performs best"

2 The Signal-plus-noise framework

2.1 The signal-plus-noise framework for a single ensemble prediction system

The signal-plus-noise representation is a widely-used and accepted statistical framework for ensemble forecasts in the climate literature.

$$x_{t,r} = \mu_x + \sigma_x(\lambda_x s_t + \sqrt{1-\lambda_x^2}\epsilon_{t,r}) \tag{1}$$

$$y_t = \mu_y + \sigma_y(\lambda_y s_t + \sqrt{1-\lambda_y^2}\epsilon'_t) \tag{2}$$

It is a believable generative joint model for ensemble members and observations, and well-studied in the statistical literature (factor analysis model, latent variable model, random effects model?). The framework models exchangeable ensemble members that are coexchangeable with the observation. Thereby, the framework can model discrepancies in mean, scale, and correlation structure. The framework has found application as a toy model in sensitivity studies, and to infer the skill of actual climate models [**?**, **?**, **?**].

In a multi-model context, the S/N framework can be applied by combining all ensembles into a single super-ensemble, whose members are judged to be exchangeable, and coexchangeable with the observations. We will address the question how to extend the S/N framework to a more flexible framework to account for differences between the individual models.

2.2 The extended S/N model for multi-model ensembles

For further discussions, it will be useful to treat the observation as a one-member ensemble. Then a decomposition of forecast and observation data into signal plus noise is written as

$$x_{t,m,r} = \mu_{x,m} + \sigma_{x,m}(\lambda_m s_t + \sqrt{1-\lambda_m^2}\epsilon_{t,m,r}) \tag{3}$$

every ensemble sees the same signal, but can have different mean, different scale, and different SNR.

A straightforward extension of the signal-plus-noise model is to assume the existence of multiple signals s_t instead of only one. That is, we represent the "systematic" part of the data as a superposition of S mutually independent signals:

$$x_{t,m,r} = \mu_{x,m} + \sigma_{x,m}\left(\sum_{j=1}^{S}\lambda_{m,j}s_{j,t} + \sqrt{1-\sum_{j=1}^{S}\lambda_{m,j}^2}\epsilon_{t,m,r}\right) \tag{4}$$

For example $S = 2$ was used in Weigel et al to represent common model error. (more on the Weigel model, also comment on triple collocation).

We next formulate the general model in matrix notation. Define the data vector $\boldsymbol{x}_t = (x_{t,1,1}, x_{t,2,1}, \ldots, x_{t,2,R_2}, \ldots x_{t,M,R_M})'$ where $x_{t,m,r}$ denotes the rth ensemble member of the mth model at time t. The first value $x_{t,1,1}$ corresponds to the observation at time t. A generative statistical model for the data vector $\boldsymbol{x}_t$ that is based on the signal-plus-noise concept is written as

$$\boldsymbol{x}_t = \boldsymbol{\mu} + \operatorname{diag}(\boldsymbol{\sigma})(\boldsymbol{\Lambda}\boldsymbol{s}_t + \operatorname{diag}(\boldsymbol{\delta})\boldsymbol{\epsilon}_t) \tag{5}$$

where

$$\begin{aligned} \boldsymbol{x}_t, \boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\delta}, \boldsymbol{\epsilon}_t &\in \mathbb{R}^{R\times 1} && (6)\\ \boldsymbol{s}_t &\in \mathbb{R}^{S\times 1} && (7)\\ \boldsymbol{\Lambda} = (\boldsymbol{\lambda}_1, \dots, \boldsymbol{\lambda}_R)' &\in \mathbb{R}^{R\times S} && (8)\\ t &= 1, \dots, N && (9) \end{aligned}$$

$\operatorname{diag}(\boldsymbol{v})$ denotes the matrix with diagonal elements equal to the elements of the vector $\boldsymbol{v}$ and zero off-diagonal elements. $R = \sum_{m=1}^{M} R_m$ is the total number of values in the data vector $\boldsymbol{x}_t$, i.e. observation and all ensemble members of the multi-model ensemble. M is the total number of models in the MME, plus one for the observation. S is the number of signals used in the framework. Furthermore, we set

$$(\boldsymbol{\delta})_i = \sqrt{1 - \boldsymbol{\lambda}_i^T \boldsymbol{\lambda}_i} \tag{10}$$

for $i = 1, \dots, R$ which ensures that the elements of $\boldsymbol{\sigma}$ represent the total variability, i.e. $\operatorname{Var}[(\boldsymbol{x}_t)_i] = (\boldsymbol{\sigma})_i^2$. This creates a constraint on the elements of $\boldsymbol{\Lambda}$: The sum of squared elements along the rows of $\boldsymbol{\Lambda}$ cannot exceed one.

Ensemble members from the same model are treated as exchangeable. Exchangeability implies that means, variances, and covariances between ensemble members within each model are equal. The assumed within-model exchangeability is modelled by setting those rows in $\boldsymbol{\mu}$, $\boldsymbol{\sigma}$, $\boldsymbol{\Lambda}$, $\boldsymbol{\delta}$ that correspond to ensemble members of the same model to be equal. Mathematically, equality is achieved by the $(0,1)$-matrix $\boldsymbol{R} \in \{0,1\}^{R\times M}$

which has exactly one 1 per row:

$$\boldsymbol{R} = \begin{pmatrix} 1 & 0 & 0 & \cdots \\ 0 & 1 & 0 & \cdots \\ \vdots & \vdots & \vdots & \\ 0 & 1 & 0 & \cdots \\ 0 & 0 & 1 & \cdots \\ \vdots & \vdots & \vdots & \end{pmatrix} \tag{11}$$

Setting $\boldsymbol{R}_{i,m} = 1$ indicates that the ith entry of $\boldsymbol{x}_t$ corresponds to the mth ensemble (or observation if $m = 1$). To give an example, assume we have 1 observation and two ensembles with 2 and 3 members, respectively. The matrix $\boldsymbol{R}$ is given by

$$\boldsymbol{R} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix} \tag{12}$$

and we can transform the reduced mean vector $\boldsymbol{\mu}^* = (\mu_1, \mu_2, \mu_3)'$ into the the block mean vector $\boldsymbol{\mu} = (\mu_1, \mu_2, \mu_2, \mu_3, \mu_3, \mu_3)'$ by multiplication:

$$\boldsymbol{\mu} = \boldsymbol{R}\boldsymbol{\mu}^* \tag{13}$$

The matrix $\boldsymbol{R}$ transforms the parameters from the model level, where parameter vectors and matrices have dimension M, to the member level, where the dimension is R. We similarly define $\boldsymbol{\Lambda}^*$, $\boldsymbol{\delta}^*$, and $\boldsymbol{\sigma}^*$ by

$$\boldsymbol{\Lambda} = \boldsymbol{R}\boldsymbol{\Lambda}^* \tag{14}$$

$$\boldsymbol{\delta} = \boldsymbol{R}\boldsymbol{\delta}^* \tag{15}$$

$$\boldsymbol{\sigma} = \boldsymbol{R}\boldsymbol{\sigma}^* \tag{16}$$

where the starred parameters describe the data on the model-level, and the non-starred parameters describe the data on the member level.

2.3 Parameter estimation

Conditional on the constant model parameters $\boldsymbol{\mu}$, $\boldsymbol{\sigma}$, and $\boldsymbol{\Lambda}$, the data vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ are independent random variables from a multivariate Normal distribution with mean vector $\boldsymbol{\mu} = \boldsymbol{R}\boldsymbol{\mu}^*$, and the $R \times R$ covariance matrix $\boldsymbol{\Sigma}$, which is given by

$$\boldsymbol{\Sigma} = \mathrm{diag}(\boldsymbol{\sigma}) \left(\boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \mathrm{diag}(\boldsymbol{\delta})^2\right) \mathrm{diag}(\boldsymbol{\sigma}) \tag{17}$$

We mentioned before that due to exchangeability of members within each model, $\boldsymbol{\Sigma}$ is highly structured – many elements of $\boldsymbol{\Sigma}$ are equal. The structure of $\boldsymbol{\Sigma}$ is expressed using the matrix $\boldsymbol{R}$, by substituting $\boldsymbol{R}\boldsymbol{\sigma}^*$ etc into eq. 17.

(Comments on number of free parameters, factor analysis, orthogonal transformations, and dimensionality reduction.)

(Compare number of free parameters between $\boldsymbol{\Sigma}$ and the $(\boldsymbol{\sigma}, \boldsymbol{\Lambda})$-parametrisation, argue that method of moments does not always provide unique solutions, comment on triple collocation)

For statistical inference by maximum likelihood estimation of Bayesian inference, we have to evaluate the Normal likelihood function, which depends on $\mathbf{\Sigma}^{-1}$ and $\log|\mathbf{\Sigma}|$. We show in the appendix how the structure of $\mathbf{\Sigma}$ can be exploited to greatly reduce the complexity of the likelihood calculation.

2.4 Post-processing: The posterior predictive distribution

The S/N model parametrises the covariance matrix of a multivariate Normal distribution for the observations and MME forecasts. We can thus work out the conditional distribution of the observation y_t, given the MME forecast $\boldsymbol{x}_t$. For fixed parameter values $\boldsymbol{\theta}$, the conditional distribution is a Normal distribution

$$y_t|x_t, \theta \sim \mathcal{N}(m_t, P_t) \tag{18}$$

with

$$m_t = \mu_y + \mathbf{\Sigma}_{1\bullet}\mathbf{\Sigma}_{\bullet\bullet}^{-1}(\boldsymbol{x}_t - \boldsymbol{\mu}_\bullet) \tag{19}$$

$$P_t = \mathbf{\Sigma}_{11} - \mathbf{\Sigma}_{1\bullet}\mathbf{\Sigma}_{\bullet\bullet}^{-1}\mathbf{\Sigma}_{\bullet 1} \tag{20}$$

Since the parameter vector $\boldsymbol{\theta}$ is not known exactly, but only its posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{X})$ is known, we have to integrate over the posterior. The posterior predictive distribution is given by

$$p(y_t|\boldsymbol{x}_t, D_{-t}) = \int d\boldsymbol{\theta}\; p(y_t|\boldsymbol{x}_t, \boldsymbol{\theta})p(\boldsymbol{\theta}|D_{-t}) \tag{21}$$

2.5 The case $S = 1$

In the case of $S = 1$, i.e., a single signal-variable, the expressions simplify and more insight can be gained into the framework.

when does the framework reduce to linear regression on the MMM

when does the framework reduce to multiple linear regression on the individual ensemble means

3 Model selection: Is there more than one predictable signal?

How should the number of signals S be chosen?

The total number of parameters in the statistical model depends on S, so S acts as a dimensionality reduction for more robust inference.

We can invoke arguments on theoretical grounds. Does the chosen framework reproduce data that conform with our beliefs about the underlying system. For example, the one-signal model implies that all models have the same correlation with the observations in the limit of infinite ensemble sizes. Do we believe this? If we think that different ensemble forecasting systems have different skill, we must choose more than one signal.

There are arguments on practical grounds. How well does the model fit the data? How does the predictive performance depend on the model specification? We might believe a multiple-signal framework to be a better model, but then find the one-signal framework to provide better predictions.

leave-one-out log predictive density, simplification in sampling-based statistical inference

4 Inference

4.1 Point estimation

method of moments

maximum likelihood

document improved performance by using the simplification

4.2 Bayesian inference

posterior probability distribution

numerical MCMC

limited benefits from using the reduced likelihood function in STAN

4.3 Prior specification

We use informative priors to include substantive knowledge about the modelled process, and to rule out a priori unreasonable parameter values.

We specify priors on observable quantities, such as ensemble mean, total variance, and correlation coefficients, and translate those into prior distributions on model parameters.

Normal prior on mean parameters

Gamma prior on variance parameters

Dirichlet priors on rows of $\mathbf{\Lambda}$ squared.

5 Application to seasonal MME forecasts of ENSO and NAO

6 Appendix

6.1 Fast calculation of the likelihood function

The S/N model assumes a joint multivariate Normal distribution of ensemble member forecasts and observations. The log-likelihood function of the multivariate normal distribution is given by

$$\begin{aligned}\ell(\boldsymbol{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\frac{N}{2}\Big[& R\log(2\pi) + \log|\boldsymbol{\Sigma}| \\ &+ \mathrm{tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{S}) + (\bar{\boldsymbol{x}} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{x}} - \boldsymbol{\mu})\Big]\end{aligned} \tag{22}$$

where $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)'$ is the data matrix, where

$$\boldsymbol{S} = \frac{1}{N}\sum_{t=1}^{N}(\boldsymbol{x}_t - \bar{\boldsymbol{x}})(\boldsymbol{x}_t - \bar{\boldsymbol{x}})' \tag{23}$$

is the sample covariance matrix, and

$$\bar{\boldsymbol{x}} = \frac{1}{N}\sum_{t=1}^{N}\boldsymbol{x}_t \tag{24}$$

is the sample mean vector. Calculating $\boldsymbol{\Sigma}^{-1}$ to evaluate the log-likelihood function requires inversion of the $R \times R$ matrix $\boldsymbol{\Sigma}$ which can be computationally expensive if the ensemble gets large. The complexity can be greatly reduced by exploiting the structure in $\boldsymbol{\Sigma}$. The inverse covariance matrix can be written as

$$\boldsymbol{\Sigma}^{-1} = \boldsymbol{R}\boldsymbol{M}\boldsymbol{R}' + \mathrm{diag}(\boldsymbol{R}\boldsymbol{\tau}^*) \tag{25}$$

where

$$(\boldsymbol{\tau}^*)_i = \frac{1}{(\boldsymbol{\delta}^*)_i^2 (\boldsymbol{\sigma}^*)_i^2} \; \forall \, i = 1, \ldots, M \tag{26}$$

and

$$\begin{aligned} \boldsymbol{M} =& -\operatorname{diag}(\boldsymbol{\sigma}^* \circ \boldsymbol{\delta}^* \circ \boldsymbol{\delta}^*)^{-1} \big[\mathbf{1}_M + \boldsymbol{\Lambda}^* \boldsymbol{\Lambda}^{*'} \\ & \boldsymbol{R}' \boldsymbol{R} \operatorname{diag}(\boldsymbol{\delta}^*)^{-2} \big]^{-1} \boldsymbol{\Lambda}^* \boldsymbol{\Lambda}^{*'} \operatorname{diag}(\boldsymbol{\sigma}^* \circ \boldsymbol{\delta}^* \circ \boldsymbol{\delta}^*)^{-1} \end{aligned} \tag{27}$$

where $\boldsymbol{v} \circ \boldsymbol{w}$ denotes the Hadamard product, i.e. element-wise multiplation, $(\boldsymbol{v} \circ \boldsymbol{w})_i = (\boldsymbol{v})_i (\boldsymbol{w})_i$ (cf. the review by Henderson and Searle 1981 on inverting sums of matrices). $\boldsymbol{\Sigma}^{-1}$ inherits the block structure of $\boldsymbol{\Sigma}$. Therefore, we have

$$\operatorname{tr}(\boldsymbol{\Sigma}^{-1} \boldsymbol{S}) = \operatorname{tr}(\boldsymbol{M} \boldsymbol{R}' \boldsymbol{S} \boldsymbol{R}) + \operatorname{tr}(\operatorname{diag}(\boldsymbol{\tau}^*) \boldsymbol{R}' \operatorname{diag}(\boldsymbol{S}) \boldsymbol{R}). \tag{28}$$

The $M \times M$ matrix $\boldsymbol{R}'\boldsymbol{S}\boldsymbol{R}$ is the matrix of block sums of the sample covariance matrix, i.e. For example, the element $(2, 3)$ of $\boldsymbol{R}'\boldsymbol{S}\boldsymbol{R}$ is the sum of covariances between all possible pairs of members from the second and third model. Likewise, $\boldsymbol{R}' \operatorname{diag}(\boldsymbol{S}) \boldsymbol{R}$ is the $(M \times M)$ matrix with diagonal elements equal to the sums of variances of the members within each model. The term $\operatorname{tr}(\boldsymbol{\Sigma}^{-1} \boldsymbol{S})$ of the likelihood function thus depends only on the within-block sums of sample variances and sample covariances.

Next, note that

$$\begin{aligned} & (\bar{\boldsymbol{x}} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\bar{\boldsymbol{x}} - \boldsymbol{\mu}) \\ =& (\boldsymbol{R}' \bar{\boldsymbol{x}} - \boldsymbol{R}' \boldsymbol{\mu})' \boldsymbol{M} (\boldsymbol{R}' \bar{\boldsymbol{x}} - \boldsymbol{R}' \boldsymbol{\mu}) \\ & + (\boldsymbol{R}' \bar{\boldsymbol{x}} - \boldsymbol{R}' \boldsymbol{\mu})' \operatorname{diag}(\boldsymbol{\tau}^*) (\boldsymbol{R}' \bar{\boldsymbol{x}} - \boldsymbol{R}' \boldsymbol{\mu}). \end{aligned} \tag{29}$$

The previous equations show that the sufficient statistics of the S/N model are the within-model block sums of covariances, variances, and means.

Using the Matrix Determinant Lemma, we can simplify $\log|\boldsymbol{\Sigma}|$ as follows.

$$\begin{aligned} \log|\boldsymbol{\Sigma}| &= \log|\mathbf{1}_S + \boldsymbol{\Lambda}'\boldsymbol{R}'\operatorname{diag}(\boldsymbol{\delta})^{-2}\boldsymbol{R}\boldsymbol{\Lambda}| && (30) \\ &\quad + 2\sum_{m=1}^{M} R_m \log[(\boldsymbol{\delta}^*)_m(\boldsymbol{\sigma}^*)_m] && (31) \end{aligned}$$