Journal of Machine Learning Research 22 (2021) 1-40 Submitted 1/21; Revised 6/21; Published 7/21

Improved Shrinkage Prediction under a Spiked Covariance Structure

Trambak Banerjee TRAMBAK@KU.EDU
Analytics, Information and Operations Management
University of Kansas
Lawrence, KS 66045, USA

Gourab Mukherjee GOURAB@USC.EDU
Data Sciences and Operations
University of Southern California
Los Angeles, CA 90089, USA

Debashis Paul DEBPAUL@UCDAVIS.EDU
Department of Statistics
University of California, Davis
Davis, CA 95616, USA

Editor: Daniela Witten

Abstract

We develop a novel shrinkage rule for prediction in a high-dimensional non-exchangeable hierarchical Gaussian model with an unknown spiked covariance structure. We propose a family of priors for the mean parameter, governed by a power hyper-parameter, which encompasses independent to highly dependent scenarios. Corresponding to popular loss functions such as quadratic, generalized absolute, and Linex losses, these prior models induce a wide class of shrinkage predictors that involve quadratic forms of smooth functions of the unknown covariance. By using uniformly consistent estimators of these quadratic forms, we propose an efficient procedure for evaluating these predictors which outperforms factor model based direct plug-in approaches. We further improve our predictors by considering possible reduction in their variability through a novel coordinate-wise shrinkage policy that only uses covariance level information and can be adaptively tuned using the sample eigen structure. Finally, we extend our disaggregate model based methodology to prediction in aggregate models. We propose an easy-to-implement functional substitution method for predicting linearly aggregated targets and establish asymptotic optimality of our proposed procedure. We present simulation experiments as well as real data examples illustrating the efficacy of the proposed method.

Keywords: Shrinkage predictors, Asymptotic optimality, Spiked covariance, Non-exchangeable priors, Aggregated predictors, Asymmetric loss, Generalized absolute loss, Linex loss.

1. Introduction

In every branch of big data analytics, it is now commonplace to use notions of shrinkage for the construction of robust algorithms and predictors. Over the last decade, driven by applications in a wide range of scientific problems, the traditional roles of statistical shrinkage have rapidly evolved as new perspectives have been introduced to address and exploit complex, latent structural properties of modern datasets. Incorporating such structural properties vastly improves predictive efficiency. Traditional shrinkage estimators in high-dimensional location models (see Efron (2012); Fourdrinier, Strawderman, and Wells (2017); Zhang (2003); Robbins (1985); Greenshtein and Ritov (2009); Greenshtein and Park (2009); Koenker and Mizera (2014); Dicker and Zhao (2016); Efron and Hastie (2016); Brown and Greenshtein (2009) and the references therein) were developed based on homoscedastic models using notions of spherical symmetry. Recent results of Xie, Kou, and Brown (2012, 2016); Weinstein, Ma, Brown, and Zhang (2018); Tan (2015) and Brown, Mukherjee, and Weinstein (2018) have brought to light new shrinkage phenomena in heteroscedastic models. However, these results are based on multivariate setups with known covariances. In a host of modern applications which are briefly described below, we need simultaneous predictions of several dependent variables when the covariance is unknown and needs to be estimated. Here, we develop efficient shrinkage predictors in these high-dimensional problems with an unknown covariance.

1. In financial portfolio selection, the vector of next period excess returns on investable assets form a critical component in determining optimal portfolio weights (Karoui et al., 2011). Different prediction programs are employed to estimate the future returns with several popular approaches using factor covariance models (Fan, Fan, and Lv, 2008; Johnstone and Titterington, 2009) to capture the dependence among asset returns (Kozak et al., 2020).

2. In cell biology, the problems of predicting the expressions of several genes leads to inference in a high-dimensional location model (Cavrois et al., 2017; Sen et al., 2014). Effective statistical methods usually integrate the dependence structure of gene expressions while conducting inference on such high-dimensional location parameters (Sun and Cai, 2009).

3. In health care management, simultaneous prediction of several inventories or resources is very important for optimal operations (Green et al., 2013). The loss function for the health care provider is agglomerative across different hospitals and the individual losses are asymmetric as a hospital would incur a shortfall cost if too many patients arrive and an excess capacity cost if too many regular nurses are scheduled compared to the number of patients. In this paper, we study shrinkage prediction under such loss functions. Mukherjee et al. (2015) showed that in uncorrelated models, empirical Bayes induced shrinkage can provide better performance than simple coordinate-wise rules. Incorporating the dependence structure among the patient arrivals in different hospitals would improve shrinkage rules.

4. A topic of current research in supply chain management (Ban and Rudin, 2019) is the inventory optimization problem of distributors and retailers who, based on past sales

data, need to predict future demands and balance the trade-offs between incurring high depreciation costs on unsold inventory versus suffering lost sales costs. Here, we study the optimal stocking problem by analyzing grocery sales data across several retail outlets in the USA. For any distributor forecasting future sales across so many outlets, this translates to a high-dimensional demand prediction problem where incorporating co-dependencies in the demands among different stores is potentially useful.

We propose CASP – a *Coordinate-wise Adaptive Shrinkage Prediction* rule for shrinkage prediction in high-dimensional Gaussian models with an unknown mean and covariance. We consider a hierarchical framework that incorporates a new family of non-exchangeable priors. Based on this, we develop shrinkage predictive rules for several popular loss functions including symmetric as well as asymmetric losses. Traditional shrinkage algorithms in the existing literature are developed under exchangeable priors (Fourdrinier et al., 2017). Our non-exchangeable hierarchical framework connects unknown location and covariance parameters. This can be seen as a means of controlling the complexity of the model, since the Bayes predictive rules involve the covariance. Predictive difficulty in this setup depends on the amount of available auxiliary information about the unknown covariance. When such information is limited, structural assumptions about the covariance are needed to improve the accuracy of the predictive rules. Here, motivated by applications, we assume a low dimensional spiked covariance (factor model) structure akin to Johnstone and Paul (2018).

In our novel hierarchical setup, we derive the Bayes predictive rule under different types of loss functions. They involve quadratic forms in functionals of the unknown population covariance. Thus, the predictive rules need to be accurately estimated by using available information about the covariance. Plug-in approaches that replace the unknown covariance by its unbiased estimator are not optimal. Our proposed method is based on uniform estimation of quadratic forms involving the unknown covariance and is proved to be asymptotically optimal. This quadratic form estimation step is a key ingredient in our procedure. It leverages the spiked covariance structure and includes corrections based on the phase transition phenomena associated with high-dimensional principal component analysis (Ma, 2013; Cai et al., 2013). Additionally, to obtain better bias-variance trade-offs in terms of the estimated prediction rule, we introduce a novel coordinate-wise shrinkage policy that ties in nicely with the structure of the loss function.

Finally, we extend our methodology for prediction to aggregated models. In many contemporary applications, the interest is not in predicting high-dimensional future observation vectors but in predicting prespecified linear functionals of future observations. Prediction in aggregated models is more complicated than prediction in disaggregate models. Using a substitution rule, we establish that our proposed methodology can also work in aggregated models. This, we believe, greatly enhances the scope and applicability of the proposed methods over those in the existing literature.

The paper is organized as follows. In Section 2, we describe our predictive setup. In Section 3, our proposed methodology CASP and its asymptotic properties are presented. In Section 4, we further develop the CASP method for prediction in aggregated models. Numerical performances of our methods are investigated using both simulated and real data in Sections 5 and 6, respectively. Proofs and additional technical details are relegated to the Appendix.

2. Predictive model

Following the statistical prediction analysis framework of Aitchison and Dunsmore (1976) and Geisser (1993), consider the high-dimensional prediction setup where the observed past $\boldsymbol{X} = (X_1, \ldots, X_n)$ as well as the future observation $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ are distributed according to a normal distribution with an unknown mean $\boldsymbol{\theta}$ and an unknown covariance proportional to $\boldsymbol{\Sigma}$. The past and the future are related only through the unknown parameters $\boldsymbol{\theta}$ and $\boldsymbol{\Sigma}$, conditioned on which they are independent. Thus, the n-dimensional predictive model has:

$$\texttt{Past observations } \boldsymbol{X} \sim N_n(\boldsymbol{\theta}, \boldsymbol{\Sigma}), \texttt{ and Future } \boldsymbol{Y} \sim N_n(\boldsymbol{\theta}, m_0^{-1}\boldsymbol{\Sigma}) \ , \tag{1}$$

where $\boldsymbol{\Sigma}$ is an unknown $n \times n$ positive definite matrix, $\boldsymbol{\theta}$ is an unknown $n \times 1$ vector and $m_0 > 0$ is a known constant (typically $m_0 = 1$). The setup in equation (1) has been studied in George et al. (2006), George and Xu (2008), and Mukherjee et al. (2015), with known diagonal covariances. As in Xie et al. (2012) and Kou and Yang (2017), for combining information across dimensions and achieving pooling based enhanced prediction, we consider an hierarchical model (Dey et al., 2000; Gelman and Hill, 2006; Banerjee et al., 2014) that puts a conditional prior on the location $\boldsymbol{\theta}$. In many contemporary applications involving correlated Gaussian models, non-exchangeable priors are needed to suitably incorporate covariance information in the hierarchical framework (Pástor and Stambaugh, 2000; Pástor, 2000; Harvey et al., 2016).

Here, we impose a class of conjugate priors on the location parameter $\boldsymbol{\theta}$. The prior family is governed by hyper-parameters $\boldsymbol{\eta}$, τ and β, and involves power-decay on the unknown covariance:

$$\pi(\boldsymbol{\theta}|\boldsymbol{\Sigma}, \boldsymbol{\eta}, \tau, \beta) \sim N_n\Big(\boldsymbol{\eta},\ \tau\, \boldsymbol{\Sigma}^{\beta}\Big). \tag{2}$$

The *shape* parameter β is key to controlling the essential characteristic of the posterior density of $\boldsymbol{\theta}$ under model (1). As β varies in $[0, \infty)$, it produces a large family of priors capable of reflecting perfect independence to highly dependent scenarios. When $\beta = 0$, the exchangeable prior on the locations resembles the setup of Xie et al. (2012) with known diagonal covariance. With $\beta = 1$, the prior has the same correlation structure as the data, whereas with $\beta > 1$, the prior is relatively less concentrated towards the dominant variability directions in the data. In the finance literature, this family of priors is widely used in asset pricing for formulating varied economically motivated priors that induce shrinkage estimation of market factors (Kozak et al., 2020). While $\beta = 0$ corresponds to the diffuse prior in Harvey et al. (2016), $\beta = 1$ gives the asset pricing prior in Pástor and Stambaugh (2000); Pástor (2000) and $\beta = 2$ yields the prior proposed in Kozak et al. (2020) that shrinks the contributions of low-variance principal components of the candidate factors. The *scale* parameter τ is allowed to vary between 0 to ∞. The location parameter $\boldsymbol{\eta}$ is usually restricted to some prespecified low dimensional subspace.

2.1 Aggregated prediction objectives

In the hierarchical setup of equations (1), (2) and based on observing $\boldsymbol{x}$, our goal is to predict $\boldsymbol{Y}$ by $\hat{\boldsymbol{q}}(\boldsymbol{x})$ under a loss that is additive across the n dimensions. Along with the

problem of predicting $\boldsymbol{Y}$, we also consider predicting several linearly aggregated components in model (1). The predictor here is $\boldsymbol{V} = \boldsymbol{AY}$, where the transformation matrix $\boldsymbol{A} \in \mathbb{R}^{p\times n}$ is observed with $p \leq n$ and full rank. Instead of predicting the raw $\boldsymbol{Y}$ at the disaggregate level, the goal is to formulate $\hat{\boldsymbol{q}} = \{\hat{q}_i(\boldsymbol{X}) : 1 \leq i \leq p\}$ based on the disaggregated past data $\boldsymbol{X}$ such that $\hat{\boldsymbol{q}}$ optimally forecasts the aggregated predictor $\boldsymbol{V}$. The loss function is again cumulative but across the p components of $\boldsymbol{V}$.

Aggregate prediction problems often arise in several applications. For example, in portfolio selection (Pástor, 2000), $\boldsymbol{A}$ represents the $p \times n$ portfolio weight matrix of p investors and $\boldsymbol{Y}$ is the next period excess return vector on the n assets. Similarly, as discussed in Section 6, in supply chain management distributors need to forecast the future sales of their products across a large number of retail outlets spread over various locations or states. Often, the high inter-state transfer costs forbid the distributors from delivering their products to these retail outlets from a central warehouse. Instead, the products are typically sourced at regional or state warehouses which are then distributed to the retail outlets in the nearby region. Thus, in this demand forecasting setup, the matrix $\boldsymbol{A}$ might represent the $p \times n$ aggregation matrix that aggregates the demand for each product across the n retail outlets into p states, and $\boldsymbol{Y}$ is the n dimensional future demand vector at each retail outlet. Such problems, where the target distribution is different from that of past observations, are more challenging than disaggregate level prediction where the target is $\boldsymbol{Y}$ itself (Komaki, 2015; Yano and Komaki, 2017; George and Xu, 2008). Naturally, when $p = n$ and $\boldsymbol{A} = \boldsymbol{I}_n$, we revert to prediction at the disaggregate level.

2.2 Loss functions

Prediction accuracy depends on the loss function used, even more so than in an estimation problem. We consider three popular loss functions that routinely arise in applications: quadratic loss, generalized absolute loss, and Linex loss. While quadratic loss is the most widely studied loss function in Statistics, the utility and necessity of asymmetric losses, like the generalized absolute loss and Linex loss, has long been acknowledged, for instance in the works of Koenker and Bassett Jr (1978), Zellner (1986). We next discuss the two aforementioned asymmetric losses and then present the Bayes predictors in our hierarchical setup for these losses.

Generalized absolute loss function, also referred to as the check loss (see Chapter 11.2.3 of Press (2009)), is a piecewise linear loss function with two linear segments and uses differential linear weights to measure the amount of underestimation or overestimation. It is the simplest as well as the most popular asymmetric loss function and is fundamental in quantile regression (Koenker and Bassett Jr, 1978). If $\hat{q}_i(\boldsymbol{X})$ represents the predictive estimate of the future V_i, then under generalized absolute loss the i^{th} coordinate incurs a loss

$$\mathcal{L}_i(V_i, \hat{q}_i(\boldsymbol{x})) = b_i(V_i - \hat{q}_i)^+ + h_i(\hat{q}_i - V_i)^+ \tag{3}$$

where b_i, h_i are known positive costs associated with underestimation and overestimation, respectively, in coordinate i, and $a^+ \coloneqq \max(a, 0)$. In inventory management problems (Mukherjee et al., 2015; Ban and Rudin, 2019) for example, where overestimation leads to depreciation and storage costs, but underestimation may lead to significant reputation costs for the retailers, the generalized absolute loss function arises naturally with $b_i \gg h_i$. When

$b_i = h_i$, equation (3) represents the usual ℓ_1 loss for coordinate i.

Linex loss function (Varian, 1975), on the other hand, uses a combination of linear and exponential functions (hence its name) to measure errors in the two different directions. The loss associated with coordinate i is

$$\mathcal{L}_i(V_i, \hat{q}_i(\boldsymbol{x})) = b_i\Big\{e^{a_i(\hat{q}_i - V_i)} - a_i(\hat{q}_i - V_i) - 1\Big\} \tag{4}$$

where $a_i \neq 0, b_i > 0$ for all i. This loss function is more appropriate for event analysis such as predicting accident counts or crime rates, underestimations of which result in much graver consequences than overestimations. However, for small values of $|a_i|$, the Linex loss behaves approximately like a quadratic loss function (Zellner, 1986).

2.3 Bayes predictors under known covariance

We define a few notations first. Let $l_p(\boldsymbol{V}, \hat{\boldsymbol{q}}) = p^{-1}\sum_{i=1}^{p}\mathcal{L}_i(V_i, \hat{q}_i)$ denote the average loss for predicting $\boldsymbol{V}$ using $\hat{\boldsymbol{q}}$ which depends on $\boldsymbol{X}$. For each $\boldsymbol{X} = \boldsymbol{x}$, the associated *predictive loss* is $L_p(\boldsymbol{\psi}, \hat{\boldsymbol{q}}) = \mathbb{E}_{\boldsymbol{V}} l_p(\boldsymbol{V}, \hat{\boldsymbol{q}})$ where $\boldsymbol{\psi} = \boldsymbol{A\theta}$ and the expectation is taken over the distribution of the future $\boldsymbol{V}$ only. The *predictive risk* is given by $\mathbb{E}_{\boldsymbol{X}} L_p(\boldsymbol{\psi}, \hat{\boldsymbol{q}})$ which, by sufficiency, reduces to $R_p(\boldsymbol{\psi}, \hat{\boldsymbol{q}}) = \mathbb{E}_{\boldsymbol{AX}} L_p(\boldsymbol{\psi}, \hat{\boldsymbol{q}})$ wherein the expectation is taken over the distribution of $\boldsymbol{AX}$. Note that the expectation over $\boldsymbol{V}$ is already included in L_p; also, the dependence of the risk on $\boldsymbol{\Sigma}$ is not made notationally explicit for ease of presentation. With $\check{\boldsymbol{\Sigma}}_\beta = \boldsymbol{A\Sigma}^\beta\boldsymbol{A}^T$, define

$$G_{r,\alpha,\beta} \coloneqq G_{r,\alpha,\beta}(\boldsymbol{\Sigma}, \boldsymbol{A}) = (\check{\boldsymbol{\Sigma}}_1^{-1} + \tau^{-1}\check{\boldsymbol{\Sigma}}_\beta^{-1})^{-r}\check{\boldsymbol{\Sigma}}_1^{\alpha},$$

where the dependence of $G_{r,\alpha,\beta}$ on τ has been kept implicit for notational ease. Our goal is to minimize $R_p(\boldsymbol{\psi}, \hat{\boldsymbol{q}})$ over the class of estimators $\hat{\boldsymbol{q}}$ for all values of $\boldsymbol{\psi}$. An essential intermediate quantity in that direction is the Bayes predictive rule $\boldsymbol{q}^{\mathsf{Bayes}}$, which is the unique minimizer of the integrated Bayes risk $B_p(\tau, \beta) = \int R_p(\boldsymbol{\psi}, \hat{\boldsymbol{q}})\pi(\boldsymbol{\psi}|\boldsymbol{\Sigma}, \tau, \beta)d\boldsymbol{\psi}$. Set $\boldsymbol{\eta}$ to a prespecified value $\boldsymbol{\eta}_0$, then Lemma 1 below provides the univariate Bayes estimator q_i^{Bayes} for the loss functions discussed earlier.

Lemma 1 (Univariate Bayes Estimator)**.** *Consider the hierarchical model in equations* (1) *and* (2)*. If $\boldsymbol{\Sigma}$ were known, the unique minimizer of the integrated Bayes risk for coordinate i is*

$$\boldsymbol{q}_i^{\mathsf{Bayes}}(\boldsymbol{AX}|\boldsymbol{\Sigma}, \tau, \beta) = \boldsymbol{e}_i^T\boldsymbol{A\eta}_0 + \boldsymbol{e}_i^T G_{1,-1,\beta}\boldsymbol{A}(\boldsymbol{X} - \boldsymbol{\eta}_0) + \mathcal{F}_i^{\mathsf{loss}}(\boldsymbol{\Sigma}, \boldsymbol{A}, \tau, \beta), \tag{5}$$

where $\boldsymbol{e}_i$ is the canonical basis vector with 1 at the i^{th} coordinate and $i = 1, \ldots, p$. With $\tilde{b}_i = b_i/(b_i + h_i)$, we have $\mathcal{F}_i^{\mathsf{loss}}(\boldsymbol{\Sigma}, \boldsymbol{A}, \tau, \beta)$ as follows:

$$\mathcal{F}_i^{\mathsf{loss}}(\boldsymbol{\Sigma}, \boldsymbol{A}, \tau, \beta) = \begin{cases} \Phi^{-1}(\tilde{b}_i)\Big(\boldsymbol{e}_i^T G_{1,0,\beta}\boldsymbol{e}_i + m_0^{-1}\boldsymbol{e}_i^T G_{0,1,0}\boldsymbol{e}_i\Big)^{1/2}, & \textit{for generalized absolute loss} \\ -\dfrac{a_i}{2}\Big(\boldsymbol{e}_i^T G_{1,0,\beta}\boldsymbol{e}_i + m_0^{-1}\boldsymbol{e}_i^T G_{0,1,0}\boldsymbol{e}_i\Big), & \textit{for Linex loss} \\ 0, & \textit{for quadratic loss.} \end{cases}$$

As τ and β vary, $\boldsymbol{q}_i^{\mathsf{Bayes}}(\boldsymbol{AX}|\boldsymbol{\Sigma}, \tau, \beta)$ constitutes a class of shrinkage predictors. When $\boldsymbol{\Sigma}$ is unknown, perhaps the simplest approach will be to plug in an estimate $\hat{\boldsymbol{\Sigma}}$ in $G_{r,\alpha,\beta}$. However, this plug-in approach produces a biased, suboptimal predictor. Next, we discuss prediction under an unknown covariance.

2.4 Prediction under an unknown covariance and structural constraints

When $\boldsymbol{\Sigma}$ is unknown, consider observing auxiliary vectors $\boldsymbol{W}_j = (W_{1j}, \ldots, W_{nj})^T$ where $\boldsymbol{W}_j|\boldsymbol{\mu}_j$ are independently distributed from $N_n(\boldsymbol{\mu}_j, \boldsymbol{\Sigma})$ for $j = 1, \ldots, m$. Here, $\boldsymbol{W} := [\boldsymbol{W}_1, \ldots, \boldsymbol{W}_m]$ and $\boldsymbol{X}$, $\boldsymbol{Y}$ from equation (1) are related only through the parameters $\boldsymbol{\mu} := \{\boldsymbol{\mu}_j : 1 \leq j \leq m\}$, $\boldsymbol{\theta}$ and $\boldsymbol{\Sigma}$. In many real world applications, $\boldsymbol{\mu_j}$s are very different from $\boldsymbol{\theta}$, and so $\boldsymbol{W}$ provides information about $\boldsymbol{\Sigma}$ but not much on $\boldsymbol{\theta}$. In these cases, the information contained in $\boldsymbol{W}$, though not useful for conducting inferences on $\boldsymbol{\theta}$, can be used to conduct inferences on $\boldsymbol{\Sigma}$. For instance, in applications involving rapid trend-changing environments with invariant correlation structures, there often exist auxiliary variables that can be used to estimate the covariances but not the average (Patton and Timmermann, 2007; Kozak et al., 2020; Harvey et al., 2016).

For a concrete example, suppose we observe $\boldsymbol{W}_t$ for $t = t_0 + 1, \ldots, t_0 + m$ time periods from the drift changing model

$$\boldsymbol{W}_t = \boldsymbol{\mu}_t + \boldsymbol{\epsilon}_t, \tag{6}$$

where $\boldsymbol{\epsilon}_t \overset{i.i.d}{\sim} N_n(\boldsymbol{0}, \boldsymbol{\Sigma})$. Here, there is a drift in the data generation process over time that does not affect the correlation structures. Let $\boldsymbol{X}$ be a vector from this model from the most recent time period t_c, i.e., $\boldsymbol{X} = \boldsymbol{\mu}_{t_c} + \boldsymbol{\epsilon}_{t_c}$ and the goal is to predict $\boldsymbol{Y} = \boldsymbol{\mu}_{t_c+1} + \boldsymbol{\epsilon}_{t_c+1}$ from the next time period $t_c + 1$. If the time lag $t_c - t_0 - m$ is large, then $\boldsymbol{\mu}_{t_c}$ is much different than $\{\boldsymbol{\mu}_t : t = t_0 + 1, \ldots, t_0 + m\}$. Most datasets available for research from the industry are of this nature and contain lagged observations, as data from the immediate past reveal current operational strategies. Note that when $t_c - t_0 - m$ and m are both large, it does not benefit us much to use $\boldsymbol{X}$ for estimating $\boldsymbol{\Sigma}$. Thus, we can concentrate on solely using $\{\boldsymbol{W}_t : t = t_0 + 1, \ldots, t_0 + m\}$ for estimating functionals involving $\boldsymbol{\Sigma}$. However, our predictions for $\boldsymbol{Y}$ will involve $\boldsymbol{X}$ as it contains pivotal information about the current drift $\boldsymbol{\mu}_{t_c+1}$. Often, $\{\boldsymbol{\mu}_t : t_0 \leq t \leq t_0 + m\}$ can be well-modeled by low-dimensional structures. In these cases, we can easily extract information regarding $\boldsymbol{\Sigma}$ from $\boldsymbol{W}$ so that the extracted information is independent of the $\boldsymbol{\mu}_t$s.

Consider using k basis functions to model $\boldsymbol{\mu}_t$ over time. Let

$$\boldsymbol{W}_t = \boldsymbol{U}\boldsymbol{C}_t + \boldsymbol{\Sigma}^{1/2}\boldsymbol{\delta}_t,$$

where $\boldsymbol{C}_t$ is a $k \times 1$ vector of basis coefficients, $\boldsymbol{U} \in \mathbb{R}^{n \times k}$ is the matrix of unknown coefficients, $\boldsymbol{\delta}_t \overset{i.i.d}{\sim} N_n(\boldsymbol{0}, \boldsymbol{I}_n)$, and $t = t_0+1, \ldots, t_0+m$. With $\boldsymbol{\mathcal{C}}_{m \times k} = (\boldsymbol{C}_{t_0+1}, \cdots, \boldsymbol{C}_{t_0+m})^T$ and the projection matrix $\boldsymbol{P}_c = \boldsymbol{\mathcal{C}}(\boldsymbol{\mathcal{C}}^T\boldsymbol{\mathcal{C}})^{-1}\boldsymbol{\mathcal{C}}^T$, note that $\boldsymbol{S} = \boldsymbol{W}(\boldsymbol{I}_m - \boldsymbol{P}_c)\boldsymbol{W}^T$ follows an n-dimensional Wishart distribution with degrees of freedom $m_k = m - k$ and mean $m_k\boldsymbol{\Sigma}$. Here, m is large and k is much smaller than m, and so we expect $m_k\boldsymbol{\Sigma}$ to be a good approximation to $m\boldsymbol{\Sigma}$. Henceforth, we assume that $\boldsymbol{S}$ follows $Wishart_n(m, \boldsymbol{\Sigma})$. Thus, $\boldsymbol{S}$ contains auxiliary information on $\boldsymbol{\Sigma}$ while $\boldsymbol{X}$ provides primary information on the unknown location $\boldsymbol{\theta}$. Moreover, the prior location $\boldsymbol{\eta}$ can be prefixed to $\boldsymbol{W}\boldsymbol{\mathcal{C}}(\boldsymbol{\mathcal{C}}^T\boldsymbol{\mathcal{C}})^{-1}C_{\mathsf{new}}$ or to the grand mean across coordinates $n^{-1}\boldsymbol{1}^T\boldsymbol{W}\boldsymbol{\mathcal{C}}(\boldsymbol{\mathcal{C}}^T\boldsymbol{\mathcal{C}})^{-1}C_{\mathsf{new}}$, where C_{new} is a known vector of basis coefficients. Note that when the covariance is unknown, predictor $\hat{\boldsymbol{q}}$ depends on $\boldsymbol{X}$ as well as $\boldsymbol{S}$. So, the *predictive risk* in this case is $\mathbb{E}_{\boldsymbol{AX},\boldsymbol{S}}L_p(\boldsymbol{\psi}, \hat{\boldsymbol{q}})$, which reduces to $R_p(\boldsymbol{\psi}, \boldsymbol{\Sigma}, \hat{\boldsymbol{q}}) = \mathbb{E}_{\boldsymbol{AX},\boldsymbol{S}}L_p(\boldsymbol{\psi}, \hat{\boldsymbol{q}})$, wherein the expectation is taken over the distribution of $\boldsymbol{AX}$ and $\boldsymbol{S}$.

If $\mathbf{\Sigma}$ is unstructured, then we need a large m for constructing good estimates of $\boldsymbol{q}_i^{\mathsf{Bayes}}$. Such a large repository of covariance information is not always available. However, there exists latent structures in $\mathbf{\Sigma}$ which can be exploited to provide efficient prediction rules for moderate m. Following Johnstone and Lu (2012), Baik and Silverstein (2006), and Paul and Aue (2014), we consider a spiked covariance structure on the unknown $\mathbf{\Sigma}$ with

$$\mathbf{\Sigma} = \sum_{j=1}^{K} \ell_j \boldsymbol{p}_j \boldsymbol{p}_j^T + \ell_0 (\boldsymbol{I}_n - \sum_{j=1}^{K} \boldsymbol{p}_j \boldsymbol{p}_j^T), \tag{7}$$

where $\boldsymbol{p}_1, \ldots, \boldsymbol{p}_K$ are orthonormal, $\ell_1 > \cdots > \ell_K > \ell_0 > 0$, and the number of spikes $1 \leq K \ll n$. While the presence of (7) in the hierarchical framework of equations (1), (2) and (6) allows efficient prediction for moderate m, directly plugging-in estimates of $\mathbf{\Sigma}$ in equation (5) will be suboptimal. By Lemma 1, it follows that accurate estimation of the prediction rule would require obtaining consistent estimates of functions $\boldsymbol{a}^T G_{r,\alpha,\beta} \boldsymbol{b}$ for arbitrary vectors $\boldsymbol{a}$ and $\boldsymbol{b}$. By using an analysis similar to Karoui and Kösters (2011), it can be shown that in regimes where $n/m \to \rho > 0$, replacing $\mathbf{\Sigma}$ by $\boldsymbol{S}$ leads to inconsistent estimation. Even using sample eigen estimates based on the spiked structure in (7) will lead to inconsistency (see Lemma 2 in the following section). To circumvent such inefficiency, we use adjusted unbiased estimates of the eigenvalues and principal eigenvectors of (7), and develop a novel methodology that can conduct uniform estimation of quadratic forms involving smooth functions of $\mathbf{\Sigma}$ that appear in equation (5).

3. Proposed methodology for disaggregated model

In Section 3.1, we first describe our proposed methodology for the efficient evaluation of the Bayes predictive rules in Lemma 1 and the asymptotic properties of the proposed methodology. Thereafter, in Section 3.2, we discuss the potential improvement in predictive efficiency that can be attained by coordinate-wise shrinkage. When $\mathbf{\Sigma}$ is unknown, evaluating (5) based on $\boldsymbol{X}$ and $\boldsymbol{S}$ reduces to estimating the quadratic forms $\boldsymbol{b}^T G_{r,\alpha,\beta} \boldsymbol{b}$ uniformly well for all τ, β where $\boldsymbol{b}$ are known vectors on the n dimensional unit sphere $\mathbb{S}_{n-1}$. In the disaggregated model $\boldsymbol{A} = \boldsymbol{I}_n$ and so, $G_{r,\alpha,\beta} = H_{r,\alpha,\beta}$ where

$$H_{r,\alpha,\beta} \coloneqq H_{r,\alpha,\beta}(\mathbf{\Sigma}) = (\mathbf{\Sigma}^{-1} + \tau^{-1}\mathbf{\Sigma}^{-\beta})^{-r}\mathbf{\Sigma}^{\alpha}.$$

When $\boldsymbol{A}$ is a general $p \times n$ rectangular matrix then $G_{r,\alpha,\beta}$ can be expressed in terms of $H_{r,\alpha,\beta}$ as follows

$$\tau^{-r} G_{r,\alpha,\beta} = \Big\{ \boldsymbol{A} H_{0,\beta,0} \boldsymbol{A}^T \Big[\boldsymbol{A} \Big(\tau H_{0,\beta,0} + H_{0,1,0} \Big) \boldsymbol{A}^T \Big]^{-1} \boldsymbol{A} H_{0,1,0} \boldsymbol{A}^T \Big\}^r \Big(\boldsymbol{A} H_{0,1,0} \boldsymbol{A}^T \Big)^{\alpha}. \tag{8}$$

We describe our procedure first for the simpler case of the disaggregated model which need estimating quadratic forms involving $H_{r,\alpha,\beta}$. Thereafter, we present the case of the aggregated model in Section 4.

3.1 Estimation of quadratic forms associated with Bayes predictors

Let the spectral decomposition of $\boldsymbol{S}$ be $\sum_{j=1}^{K} \hat{\ell}_j \hat{\boldsymbol{p}}_j \hat{\boldsymbol{p}}_j^T$ where $\hat{\boldsymbol{p}}_j$ are orthonormal and $\hat{\ell}_1 \geq \cdots \geq \hat{\ell}_n$. For unbiased estimation of the true eigenvalues and principal eigenvectors, we

adjust for the phase transition phenomenon of the eigenvalues (Baik and Silverstein, 2006; Paul, 2007; Onatski, 2012; Benaych-Georges and Nadakuditi, 2012) and eigenvectors of $\boldsymbol{S}$ (Kritchman and Nadler, 2008, 2009; Passemier and Yao, 2012; Passemier et al., 2015), when $n/m \to \rho > 0$ as $n \to \infty$ and K is fixed. As in Paul (2007) and Onatski (2012), assume the following asymptotic conditions on $\boldsymbol{\Sigma}$:

A1 Asymptotic regime : Suppose that $\rho_n = \frac{n}{m-1} \to \rho \in (0, \infty)$ as $n \to \infty$.

A2 Significant spike: Suppose that $\ell_j > \ell_0(1 + \sqrt{\rho})$ for $j = 1, \ldots, K$.

We next present unbiased estimates $\{\hat{\ell}_j^{\mathsf{e}}\}_{j=0}^K$ for the principal eigenvalues $\{\ell_j\}_{j=0}^K$ of $\boldsymbol{\Sigma}$. In what follows, if $\{R_n\}$ is a sequence of random variables and $\{a_n\}$ is a sequence of real numbers, then the notation $R_n = O_p(a_n)$ means that the sequence $\{R_n/a_n\}$ is bounded in probability while the notation $R_n = o_p(a_n)$ means that the sequence $\{R_n/a_n\}$ converges to 0 in probability as $n \to \infty$.

Define

$$\zeta(x, \rho) = \Big[\frac{1 - \rho/(x-1)^2}{1 + \rho/(x-1)}\Big]^{1/2} \text{ with } \zeta_j = \zeta(\ell_j/\ell_0, \rho).$$

Recall that under assumptions **A1** and **A2** the leading eigenvectors and eigenvalues of $\boldsymbol{S}$ have the following properties (Paul, 2007): for $j = 1, \ldots, K$,

$$\hat{\ell}_j - \ell_j \left(1 + \frac{\rho}{(\ell_j/\ell_0 - 1)}\right) = O_p(n^{-1/2}) \ , \quad j = 1, \ldots, K, \tag{9}$$

and

$$\hat{\boldsymbol{p}}_j = U_j \boldsymbol{p}_j + \sqrt{1 - U_j^2}(I - \boldsymbol{P}_K \boldsymbol{P}_K^T)\frac{\boldsymbol{\varepsilon}_j}{\sqrt{n-K}} + o_p(n^{-1/2}) \text{ with } U_j = \zeta_j + o_p(n^{-1/2}) \tag{10}$$

where $\boldsymbol{P}_K = [\boldsymbol{p}_1 : \cdots : \boldsymbol{p}_K]$, $\boldsymbol{\varepsilon}_j \sim N(0, I_{n-K})$ and the o_p term in first identity in equation (10) is in the sense of inner product. We will use these properties to ensure that the quadratic forms of the type $\boldsymbol{b}^T H_{r,\alpha,\beta} \boldsymbol{b}$ are consistently estimated. When K, the number of significant spikes, is known, we have efficient estimates $\hat{\ell}_j^{\mathsf{e}}$ of ℓ_j for $j = 0, \ldots, K$ (Passemier et al., 2017) that involve bias correction of $\hat{\ell}_j$ using the approximation properties of equation (9) as follows: Let $\hat{\ell}_0 = (n-K)^{-1} \sum_{j=K+1}^n \hat{\ell}_j$ and then for $j = 1, \ldots, K$, let $\hat{\ell}_j'$ be the solution of the following equation (for x)

$$\hat{\ell}_j = \hat{\ell}_0 \psi(x/\hat{\ell}_0, \rho_n) = x\left(1 + \frac{\rho_n}{x/\hat{\ell}_0 - 1}\right).$$

Then, the estimates of $\{\ell_j\}_{j=0}^K$ are $\{\hat{\ell}_j^{\mathsf{e}}\}_{j=0}^K$ where

$$\hat{\ell}_0^{\mathsf{e}} = \hat{\ell}_0 \left(1 + \frac{\rho_n \hat{\xi}_0}{n-K}\right) \tag{11}$$

and for $j = 1, \ldots, K$,

$$\hat{\ell}_j^{\mathsf{e}} = \frac{\hat{\ell}_0^{\mathsf{e}}}{2}\left[(\hat{\ell}_j/\hat{\ell}_0^{\mathsf{e}} + 1 - \rho_n) + \left((\hat{\ell}_j/\hat{\ell}_0^{\mathsf{e}} + 1 - \rho_n)^2 - 4\hat{\ell}_j/\hat{\ell}_0^{\mathsf{e}}\right)^{1/2}\right], \tag{12}$$

with $\hat{\xi}_0 = K + \sum_{j=1}^K \left(\hat{\ell}_j'/\hat{\ell}_0 - 1\right)^{-1}$. Now, consider the following as an estimate for $H_{r,\alpha,\beta}$:

$$\hat{H}_{r,\alpha,\beta} = \sum_{j=1}^K \frac{1}{\hat{\zeta}_j^2}(h_{r,\alpha,\beta}(\hat{\ell}_j^{\mathsf{e}}) - h_{r,\alpha,\beta}(\hat{\ell}_0^{\mathsf{e}}))\hat{\boldsymbol{p}}_j\hat{\boldsymbol{p}}_j^T + h_{r,\alpha,\beta}(\hat{\ell}_0^{\mathsf{e}})\boldsymbol{I}$$

$$= \sum_{j=1}^K \left[\frac{1}{\hat{\zeta}_j^2}h_{r,\alpha,\beta}(\hat{\ell}_j^{\mathsf{e}}) + \left(1 - \frac{1}{\hat{\zeta}_j^2}\right)h_{r,\alpha,\beta}(\hat{\ell}_0^{\mathsf{e}})\right]\hat{\boldsymbol{p}}_j\hat{\boldsymbol{p}}_j^T + h_{r,\alpha,\beta}(\hat{\ell}_0^{\mathsf{e}})\left(\boldsymbol{I} - \sum_{j=1}^K \hat{\boldsymbol{p}}_j\hat{\boldsymbol{p}}_j^T\right), \quad (13)$$

where $h_{r,\alpha,\beta}(x) = (x^{-1} + \tau^{-1}x^{-\beta})^{-r}x^{\alpha}$ is the scalar version of $H_{r,\alpha.\beta}$, $\hat{\zeta}_j = \zeta(\hat{\ell}_j^{\mathsf{e}}/\hat{\ell}_0^{\mathsf{e}}, \rho_n)$, $\hat{\ell}_0^{\mathsf{e}}$, $\hat{\ell}_j^{\mathsf{e}}$ are from equations (11), (12) respectively, and $\hat{\boldsymbol{p}}_j$ are from equation (10). A key aspect regarding the estimates $\hat{H}_{r,\alpha,\beta}$ in equation (13) is that they not only involve asymptotic adjustments to the sample eigenvalues through equations (11) and (12) but also use the phase transition phenomenon of the sample eigenvectors to appropriately adjust them through $\hat{\zeta}_j$ and equation (10).

The following condition ensures that the results on the behavior of the Bayes predictors and their estimated versions remain valid uniformly over a collection of hyper-parameters.

A3 $\tau \in \boldsymbol{T}_0$ and $\beta \in \boldsymbol{B}_0$ where $\boldsymbol{T}_0$ and $\boldsymbol{B}_0$ are compact subsets of $(0, \infty)$ and $[0, \infty)$, respectively.

Notice that **A3** implies in particular that $\tau_0 \le \tau < \infty$ for some $\tau_0 > 0$. For any fixed $c > 0$, let $\mathcal{B}$ denote a collection of $O(n^c)$ unit vectors on the n dimensional unit sphere $\mathbb{S}_{n-1}$. For the disaggregated model, Theorem 1 proves the asymptotic consistency of $\boldsymbol{b}^T\hat{H}_{r,\alpha,\beta}\boldsymbol{b}$ uniformly over the hyper-parameters (τ, β) and known vectors $\boldsymbol{b} \in \mathcal{B}$.

Theorem 1 (Asymptotic consistency of $\boldsymbol{b}^T\hat{H}_{r,\alpha,\beta}\boldsymbol{b}$)**.** *Under assumptions **A1**, **A2**, and **A3**, uniformly over $\tau \in \boldsymbol{T}_0$, $\beta \in \boldsymbol{B}_0$ and $\boldsymbol{b} \in \mathcal{B}$ with $\|\boldsymbol{b}\|_2 = 1$, we have, for all $(r, \alpha) \in \{-1, 0, 1\} \times \mathbb{R}$,*

$$\sup_{\tau \in \boldsymbol{T}_0, \beta \in \boldsymbol{B}_0, \boldsymbol{b} \in \mathcal{B}} \left|\boldsymbol{b}^T\hat{H}_{r,\alpha,\beta}\boldsymbol{b} - \boldsymbol{b}^T H_{r,\alpha,\beta}\boldsymbol{b}\right| = O_p\left(\sqrt{\frac{\log n}{n}}\right)$$

where the dependence of $H_{r,\alpha,\beta}$ on τ has been kept implicit for notational ease.

For developing the above consistent estimators, it is important to use adjusted estimates $\hat{\ell}_j^{\mathsf{e}}$ instead of the sample estimates $\hat{\ell}_j$ in equation (13). Consider the unadjusted estimator $\hat{\boldsymbol{\Sigma}}_{\mathsf{un}}$ that uses the spiked structure in (7) but unadjusted eigenvalues, and the subsequent estimator

$$\hat{H}_{r,\alpha,\beta}^{\mathsf{un}} = (\hat{\boldsymbol{\Sigma}}_{\mathsf{un}}^{-1} + \tau^{-1}\hat{\boldsymbol{\Sigma}}_{\mathsf{un}}^{-\beta})^{-r}\hat{\boldsymbol{\Sigma}}_{\mathsf{un}}^{\alpha}\ , \quad \text{where,} \quad \hat{\boldsymbol{\Sigma}}_{\mathsf{un}} = \sum_{j=1}^K \hat{\ell}_j\hat{\boldsymbol{p}}_j\hat{\boldsymbol{p}}_j^T + \hat{\ell}_0(\boldsymbol{I} - \sum_{j=1}^K \hat{\boldsymbol{p}}_j\hat{\boldsymbol{p}}_j^T).$$

The following result shows that such estimates of $\boldsymbol{b}^T H_{r,\alpha,\beta}\boldsymbol{b}$ are inefficient.

Lemma 2 (Inconsistency of spiked covariance based unadjusted estimators)**.** *Under assumptions **A1**, **A2** and **A3**, for any $\boldsymbol{b}$ with $\liminf_{n\to\infty} \max_{j=1,\ldots,K} |\boldsymbol{b}^T\boldsymbol{p}_j| > 0$ we have, $\boldsymbol{b}^T\hat{H}_{r,\alpha,\beta}^{\mathsf{un}}\boldsymbol{b} - \boldsymbol{b}^T H_{r,\alpha,\beta}\boldsymbol{b} \not\xrightarrow{p} 0$.*

An important consequence of Theorem 1 is that it allows us, almost immediately, to construct an efficient evaluation scheme for the Bayes predictive rule in Lemma 1 under the disaggregated model as follows:

Definition 1 (Predictive rule in disaggregated model)**.** *Under the hierarchical model of equations* (1) *and* (2)*, the proposed predictive rule for the disaggregated model is given by* $\hat{q}^{\mathsf{S}}$ *which is defined as*

$$\hat{q}_i^{\mathsf{S}}(\boldsymbol{X}|\boldsymbol{S},\tau,\beta) = \boldsymbol{e}_i^T\boldsymbol{\eta}_0 + \boldsymbol{e}_i^T\hat{H}_{1,-1,\beta}(\boldsymbol{X}-\boldsymbol{\eta}_0) + \hat{\mathcal{F}}_i^{\mathsf{loss}}(\boldsymbol{S},\tau,\beta) \tag{14}$$

where

$$\hat{\mathcal{F}}_i^{\mathsf{loss}}(\boldsymbol{S},\tau,\beta) = \begin{cases} \Phi^{-1}(\tilde{b}_i)\Big(\boldsymbol{e}_i^T\hat{H}_{1,0,\beta}\boldsymbol{e}_i + m_0^{-1}\boldsymbol{e}_i^T\hat{H}_{0,1,0}\boldsymbol{e}_i\Big)^{1/2}, & \textit{for generalized absolute loss} \\ -\dfrac{a_i}{2}\Big(\boldsymbol{e}_i^T\hat{H}_{1,0,\beta}\boldsymbol{e}_i + m_0^{-1}\boldsymbol{e}_i^T\hat{H}_{0,1,0}\boldsymbol{e}_i\Big), & \textit{for Linex loss} \\ 0, & \textit{for quadratic loss} \end{cases}$$

and $\tilde{b}_i = b_i/(b_i+h_i)$.

Note that $\hat{\boldsymbol{q}}^{\mathsf{S}}$ is a shrinkage rule that estimates the Bayes predictive rules in Lemma 1 under the disaggregated model by replacing the quadratic forms $\boldsymbol{b}^T H_{r,\alpha,\beta}\boldsymbol{b}$ by their consistent estimates $\boldsymbol{b}^T\hat{H}_{r,\alpha,\beta}\boldsymbol{b}$ from equation (13). Note that the second term in equation (14), which is of the form $\boldsymbol{b}^T H_{r,\alpha,\beta}\boldsymbol{c}$ (where $\boldsymbol{b}$, $\boldsymbol{c}$ are unit vectors), can also be written as a difference of two quadratic forms $(1/4)\{(\boldsymbol{b}+\boldsymbol{c})^T H_{r,\alpha,\beta}(\boldsymbol{b}+\boldsymbol{c}) - (\boldsymbol{b}-\boldsymbol{c})^T H_{r,\alpha,\beta}(\boldsymbol{b}-\boldsymbol{c})\}$, and Theorem 1 can be directly applied to yield Lemma 3 which provides decision theoretic guarantees on the predictors. It shows that uniformly over (τ,β) the largest coordinate-wise gap between $\hat{\boldsymbol{q}}^{\mathsf{S}}$ and $\boldsymbol{q}^{\mathsf{Bayes}}$ is asymptotically small.

Lemma 3. *Under assumptions* ***A1****,* ***A2*** *and* ***A3****, uniformly over* $\tau \in \boldsymbol{T}_0$*,* $\beta \in \boldsymbol{B}_0$*, for all* $(r,\alpha) \in \{-1,0,1\}\times\mathbb{R}$*, we have, conditionally on* $\boldsymbol{X}$*,*

$$\frac{\sup_{\tau\in\boldsymbol{T}_0,\beta\in\boldsymbol{B}_0}\Big\|\hat{\boldsymbol{q}}^{\mathsf{S}}(\boldsymbol{X}|\boldsymbol{S},\tau,\beta) - \boldsymbol{q}^{\mathsf{Bayes}}(\boldsymbol{X}|\boldsymbol{\Sigma},\tau,\beta)\Big\|_\infty}{\Big\|\boldsymbol{X}-\boldsymbol{\eta}_0\Big\|_2 \vee 1} = O_p\Big(\sqrt{\frac{\log n}{n}}\Big)\ .$$

While $\hat{\boldsymbol{q}}^{\mathsf{S}}$ is an asymptotically unbiased approximation to $\boldsymbol{q}^{\mathsf{Bayes}}$, the average L_2 distance between $\hat{\boldsymbol{q}}^{\mathsf{S}}$ and $\boldsymbol{q}^{\mathsf{Bayes}}$ is a non-trivial quantity due to the intrinsic variability in $\boldsymbol{X}$. In the following subsection, we introduce our Coordinate-wise Adaptive Shrinkage Prediction rule, CASP, that relies on data-driven adaptive shrinkage factors to reduce the marginal variability of $\hat{\boldsymbol{q}}^{\mathsf{S}}$ for any fixed $\boldsymbol{S}$, and minimizes the squared error of the predictors from $\boldsymbol{q}^{\mathsf{Bayes}}$.

3.2 Improved predictive efficiency by coordinate-wise shrinkage

We introduce a class of coordinate-wise shrinkage predictive rules that includes $\hat{\boldsymbol{q}}^{\mathsf{S}}$ as a special case.

Definition 2 (Class of coordinate-wise shrinkage predictive rules)**.** *Consider a class of coordinate-wise shrinkage predictive rules* $\mathcal{Q}^{\mathsf{cs}} = \{\hat{q}_i^{cs}(\boldsymbol{X}|\boldsymbol{S}, f_i, \tau, \beta) \mid f_i \in \mathbb{R}_+, \tau \in \boldsymbol{T}_0, \beta \in \boldsymbol{B}_0\}$ *where*

$$\hat{q}_i^{\mathsf{cs}}(\boldsymbol{X}|\boldsymbol{S}, f_i, \tau, \beta) = \boldsymbol{e}_i^T \boldsymbol{\eta}_0 + f_i \boldsymbol{e}_i^T \hat{H}_{1,-1,\beta}(\boldsymbol{X} - \boldsymbol{\eta}_0) + \hat{\mathcal{F}}_i^{\mathsf{loss}}(\boldsymbol{S}, \tau, \beta),$$

with $\hat{\mathcal{F}}_i^{\mathsf{loss}}(\boldsymbol{S}, \tau, \beta)$ *as defined in Definition 1 and* $f_i \in \mathbb{R}_+$ *is a shrinkage factor depending only on* $\boldsymbol{S}$*.*

The class $\mathcal{Q}^{\mathsf{cs}}$ represents a wider class of predictive rules when compared to the linear functional form of the Bayes rule. In particular, it includes $\hat{\boldsymbol{q}}^{\mathsf{S}}$ from definition 1 when $f_i = 1$ for all i. The coordinate-wise shrinkage factors f_i do not depend on $\boldsymbol{X}$ but only on $\boldsymbol{S}$, and their role lies in reducing the marginal variability of the predictive rule as demonstrated in Lemma 4 below.

Lemma 4. *Suppose that assumptions* ***A1****,* ***A2*** *and* ***A3*** *hold. Under the hierarchical model of equations* (1) *and* (2)*, as* $n \to \infty$*,*

(a) $\mathbb{E}\Big\{\Big(\hat{q}_i^{\mathsf{cs}}(\boldsymbol{X}|\boldsymbol{S}, f_i, \tau, \beta) - q_i^{\mathsf{Bayes}}(\boldsymbol{X}|\boldsymbol{\Sigma}, \tau, \beta)\Big)^2\Big\}$ *is minimized at*

$$f_i^{\mathsf{OR}} = \frac{\boldsymbol{e}_i^T U(\boldsymbol{\Sigma}) \boldsymbol{e}_i}{\boldsymbol{e}_i^T \hat{H}_{1,-1,\beta} \mathcal{J}(\boldsymbol{\Sigma}) \hat{H}_{1,-1,\beta} \boldsymbol{e}_i} + O_p\Big(\sqrt{\frac{\log n}{n}}\Big),$$

where $U(\boldsymbol{\Sigma}) \coloneqq H_{1,-1,\beta} \mathcal{J}(\boldsymbol{\Sigma}) H_{1,-1,\beta}$*,* $\mathcal{J}(\boldsymbol{\Sigma}) \coloneqq \boldsymbol{\Sigma} + \tau \boldsymbol{\Sigma}^\beta$ *and the expectation is taken with respect to the marginal distribution of* $\boldsymbol{X}$ *with* $\boldsymbol{S}$ *fixed.*

(b) For any fixed τ*,* β*, with probability 1,*

$$\limsup_{n \to \infty} \max_{1 \le i \le n} f_i^{\mathsf{OR}} \le 1.$$

Moreover, let $\mathcal{M} = \{1 \le i \le n : \|\boldsymbol{P}_K \boldsymbol{e}_i\|_2 > 0\}$*, where* $\boldsymbol{P}_K$ *denotes the* K*-dimensional projection matrix associated with the* K *spiked eigenvalues of* $\boldsymbol{\Sigma}$*. Then, with* $j(x) := x + \tau x^\beta$ *as the scalar version of* $\mathcal{J}(\boldsymbol{\Sigma})$*, we have*

$$\max_{i \in \mathcal{M}} f_i^{\mathsf{OR}} \le \max_{i \in \mathcal{M}} \frac{\boldsymbol{e}_i^T U(\boldsymbol{\Sigma}) \boldsymbol{e}_i}{\boldsymbol{e}_i^T U(\boldsymbol{\Sigma}) \boldsymbol{e}_i + j(\ell_0)\big\{h_{1,-1,\beta}(\ell_K) - h_{1,-1,\beta}(\ell_0)\big\}^2 \|\boldsymbol{P}_K \boldsymbol{e}_i\|_2^2} + O_p\left(\sqrt{\frac{\log n}{n}}\right),$$

so that the leading term on the right hand side is less than 1.

(c) Also, for any fixed τ *and* β*, we have with probability 1:*

$$\liminf_{n \to \infty} \frac{\mathbb{E}\|\hat{\boldsymbol{q}}^{\mathsf{S}}(\boldsymbol{X}|\boldsymbol{S}, \tau, \beta) - \boldsymbol{q}^{\mathsf{Bayes}}(\boldsymbol{X}|\boldsymbol{\Sigma}, \tau, \beta)\|_2^2}{\mathbb{E}\|\hat{\boldsymbol{q}}^{\mathsf{cs}}(\boldsymbol{X}|\boldsymbol{S}, \boldsymbol{f}^{\mathsf{OR}}, \tau, \beta) - \boldsymbol{q}^{\mathsf{Bayes}}(\boldsymbol{X}|\boldsymbol{\Sigma}, \tau, \beta)\|_2^2} \ge 1 \ ,$$

where the expectations are taken with respect to the marginal distribution of $\boldsymbol{X}$ *with* $\boldsymbol{S}$ *fixed.*

Lemma 4 is proved in Appendix A.7. An interesting point to note about the proof of statement (a) of the lemma is that minimizing the squared error essentially reduces to minimizing the variability of $\hat{\boldsymbol{q}}^{\mathsf{cs}}$ as any member in $\mathcal{Q}^{\mathsf{cs}}$ has asymptotically negligible bias. The optimal variance is attained by the oracle shrinkage factors f_i^{OR} which assume the knowledge of $H_{1,-1,\beta}$ and $\mathcal{J}(\boldsymbol{\Sigma})$. Statement (b) shows that these shrinkage factors lie in $[0,1]$. It also shows that some of them are actually quite different from 1. Thus, the resulting coordinate-wise shrunken oracle prediction rule greatly differs from $\hat{\boldsymbol{q}}^{\mathsf{S}}$. Indeed, statement (b) shows that if the eigenvectors of $\boldsymbol{\Sigma}$ are relatively sparse, so that for a small number of coordinates i, the quantities $\|\boldsymbol{P}_K\boldsymbol{e}_i\|_2$ are positive (and relatively large), then the shrinkage factor f_i^{OR} for the corresponding coordinates can be significantly smaller than 1. Statement (c) trivially follows from (b) and guarantees that $\hat{\boldsymbol{q}}^{\mathsf{cs}}$ constructed based on the oracle shrinkage factors f_i^{OR} are at least as good as $\hat{\boldsymbol{q}}^{\mathsf{S}}$ in terms of squared error distance from the true $\boldsymbol{q}^{\mathsf{Bayes}}$ predictor. However, as $\boldsymbol{\Sigma}$ is unknown, f_i^{OR} cannot be computed in practice. Theorem 1 allows us to estimate the oracle shrinkage factors consistently and those estimates form a key ingredient in our proposed predictive rule CASP in Definition 3 below.

Definition 3 (CASP)**.** *The coordinate-wise adaptive shrinkage prediction rule is given by* $\hat{\boldsymbol{q}}^{\mathsf{casp}} \in \mathcal{Q}^{\mathsf{cs}}$ *with* $f_i = \hat{f}_i^{\mathsf{prop}}$ *where*

$$\hat{f}_i^{\mathsf{prop}} = \frac{\boldsymbol{e}_i^T \tau \hat{H}_{1,\beta-1,\beta}\boldsymbol{e}_i}{\boldsymbol{e}_i^T \hat{R}\boldsymbol{e}_i},$$

and

$$\hat{R} = \tau\hat{H}_{1,\beta-1,\beta} + j(\hat{\ell}_0^{\mathsf{e}})\sum_{j=1}^{K}\hat{\zeta}_j^{-4}\Big\{h_{1,-1,\beta}(\hat{\ell}_j^{\mathsf{e}}) - h_{1,-1,\beta}(\hat{\ell}_0^{\mathsf{e}})\Big\}^2\hat{\boldsymbol{p}}_j\hat{\boldsymbol{p}}_j^T,$$

with $j(x) \coloneqq x + \tau x^\beta$ *as the scalar version of* $\mathcal{J}(\boldsymbol{\Sigma})$.

Unlike the numerator, the denominator in f_i^{OR} is not linear in $H_{r,\alpha,\beta}$ and estimating it with desired precision involves second order terms in $\hat{R}$. Lemma 5 below shows that indeed $\hat{f}_i^{\mathsf{prop}}$ is a consistent estimator of f_i^{OR} under our hierarchical model.

Lemma 5. *Under the hierarchical model of equations* (1) *and* (2)*,*

$$\sup_{1\le i\le n}|\hat{f}_i^{\mathsf{prop}} - f_i^{\mathsf{OR}}| = O_p\Big(\sqrt{\frac{\log n}{n}}\Big)\ .$$

Using lemmata 4 (a) and 5, Theorem 2 below guarantees the oracle optimality of $\hat{\boldsymbol{q}}^{\mathsf{casp}}$ in the class $\mathcal{Q}^{\mathsf{cs}}$ in the sense that the shrinkage factors $\hat{f}_i^{\mathsf{prop}}$ reduce the squared error between CASP and the Bayes predictive rule as much as the oracle shrinkage factors f_i^{OR} would for any predictive rule in the class $\mathcal{Q}^{\mathsf{cs}}$. Proofs of Lemma 5 and Theorem 2 are provided in appendices A.7 and A.8, respectively.

Theorem 2 (Oracle optimality of CASP)**.** *Under assumptions* ***A1****,* ***A2*** *and* ***A3****, and the hierarchical model of equations* (1) *and* (2)*, we have, conditionally on* $\boldsymbol{X}$*,*

$$\sup_{\tau\in\boldsymbol{T}_0,\beta\in\boldsymbol{B}_0}\frac{\|\hat{\boldsymbol{q}}^{\mathsf{casp}}(\boldsymbol{X}|\boldsymbol{S},\hat{\boldsymbol{f}}^{\mathsf{prop}},\tau,\beta) - \hat{\boldsymbol{q}}^{\mathsf{cs}}(\boldsymbol{X}|\boldsymbol{S},\boldsymbol{f}^{\mathsf{OR}},\tau,\beta)\|_2^2}{\|\hat{\boldsymbol{q}}^{\mathsf{cs}}(\boldsymbol{X}|\boldsymbol{S},\boldsymbol{f}^{\mathsf{OR}},\tau,\beta) - \boldsymbol{e}_i^T\boldsymbol{\eta}_0\|_2^2} = O_p(\log n/n)\ .$$

3.3 Calibration of the tuning parameters

For estimating K we use the procedure described in Kritchman and Nadler (2009) that estimates K through a sequence of hypothesis tests, determining at each step whether the k^{th} sample eigenvalue came from a spike. To estimate the prior hyper-parameters (τ, β), we first note that marginally $\boldsymbol{X} \sim N_n(\boldsymbol{\eta}_0, \mathcal{J}(\boldsymbol{\Sigma}))$. Let $\mathcal{J}_{\mathsf{inv}}(\boldsymbol{\Sigma}) = (\boldsymbol{\Sigma} + \tau\boldsymbol{\Sigma}^{\beta})^{-1}$. Our scheme for choosing (τ, β) is based on an empirical Bayes approach wherein we maximize the marginal likelihood of $\boldsymbol{X}$ with respect to (τ, β) with $\mathcal{J}(\boldsymbol{\Sigma})$ and $\mathcal{J}_{\mathsf{inv}}(\boldsymbol{\Sigma})$ replaced by their estimates $\hat{\mathcal{J}} = \tau\hat{H}_{-1,1+\beta,\beta}$ and $\hat{\mathcal{J}}_{\mathsf{inv}} = \tau^{-1}\hat{H}_{1,-1-\beta,\beta}$ respectively. In particular an estimate of (τ, β) is given by

$$(\hat{\tau}, \hat{\beta}) = \underset{\tau \in \boldsymbol{T}_0, \beta \in \boldsymbol{B}_0}{\arg\max} \; -0.5 \log|\hat{\mathcal{J}}| - 0.5\big(\boldsymbol{x} - \boldsymbol{\eta}_0\big)^T \hat{\mathcal{J}}_{inv}\big(\boldsymbol{x} - \boldsymbol{\eta}_0\big). \tag{15}$$

To facilitate implementation, the maximization in equation (15) is conducted numerically over a bounded interval $[\tau_{\mathsf{lb}}, \tau_{\mathsf{ub}}] \times [\beta_{\mathsf{lb}}, \beta_{\mathsf{ub}}]$ where, in most practical applications, prior knowledge dictates the lower $(\tau_{\mathsf{lb}}, \beta_{\mathsf{lb}})$ and upper bounds $(\tau_{\mathsf{ub}}, \beta_{\mathsf{ub}})$ of the above intervals. In the simulations and real data examples of Sections 5 and 6, we use the above scheme to estimate $(\tau,\ \beta)$.

4. Methodology for the aggregated model

Recall that under the aggregated model equation (8) expresses $G_{r,\alpha,\beta}$ in terms of $H_{r,\alpha,\beta}$. To estimate $G_{r,\alpha,\beta}$ in this setting, we adopt the substitution principle and construct the following estimates of $G_{r,\alpha,\beta}$

$$\begin{aligned}
\hat{G}_{0,1,0} &= \boldsymbol{A}\hat{H}_{0,1,0}\boldsymbol{A}^T \\
\hat{G}_{1,0,\beta} &= \tau\boldsymbol{A}\hat{H}_{0,\beta,0}\boldsymbol{A}^T\Big[\boldsymbol{A}\Big(\tau\hat{H}_{0,\beta,0} + \hat{H}_{0,1,0}\Big)\boldsymbol{A}^T\Big]^{-1}\boldsymbol{A}\hat{H}_{0,1,0}\boldsymbol{A}^T \\
\hat{G}_{1,-1,\beta} &= \tau\boldsymbol{A}\hat{H}_{0,\beta,0}\boldsymbol{A}^T\Big[\boldsymbol{A}\Big(\tau\hat{H}_{0,\beta,0} + \hat{H}_{0,1,0}\Big)\boldsymbol{A}^T\Big]^{-1},
\end{aligned}$$

which appear in the functional form of CASP for aggregated models in Definition 4 below. Our main focus in the rest of this section is to show that when p, the dimension of the aggregated space, is small compared to the dimension n of the observation space, and auxiliary data are available to estimate $\boldsymbol{\Sigma}$, the simple substitution rule described above can be still very effective. Accordingly, we make the following assumption:

A4 Aggregation matrix: Suppose $p = o(n)$ and $\boldsymbol{A} \in \mathbb{R}^{p\times n}$ is such that the matrix $\boldsymbol{A}\boldsymbol{A}^T$ is invertible and has uniformly bounded condition number even as $p, n \to \infty$.

Definition 4 (CASP for aggregated models)**.** *For any fixed $\boldsymbol{A}$ obeying assumption **A4**, consider a class of coordinate-wise shrinkage predictive rules $\mathcal{Q}^{\mathsf{cs}}_{\boldsymbol{A}} = \{\hat{q}^{cs}_i(\boldsymbol{A}\boldsymbol{X}|\boldsymbol{S}, f_i, \tau, \beta) \mid f_i \in \mathbb{R}_+, \tau \in \boldsymbol{T}_0, \beta \in \boldsymbol{B}_0\}$ where*

$$\hat{q}^{\mathsf{cs}}_i(\boldsymbol{A}\boldsymbol{X}|\boldsymbol{S}, f_i, \tau, \beta) = \boldsymbol{e}_i^T\boldsymbol{A}\boldsymbol{\eta}_0 + f_i\boldsymbol{e}_i^T\hat{G}_{1,-1,\beta}\boldsymbol{A}(\boldsymbol{X} - \boldsymbol{\eta}_0) + \hat{\mathcal{F}}^{\mathsf{loss}}_i(\boldsymbol{S}, \boldsymbol{A}, \tau, \beta),$$

and $\hat{\mathcal{F}}^{\mathsf{loss}}_i(\boldsymbol{S}, \boldsymbol{A}, \tau, \beta)$ are the estimates of $\mathcal{F}^{\mathsf{loss}}_i(\boldsymbol{\Sigma}, \boldsymbol{A}, \tau, \beta)$ as defined in Lemma 1 with $G_{r,\alpha,\beta}$ replaced by $\hat{G}_{r,\alpha,\beta}$ and $f_i \in \mathbb{R}_+$ are shrinkage factors depending only on $\boldsymbol{S}$ and $\boldsymbol{A}$. The

coordinate-wise adaptive shrinkage predictive rule for the aggregated model is given by $\hat{\boldsymbol{q}}^{\mathsf{casp}} \in \mathcal{Q}^{\mathsf{cs}}_{\boldsymbol{A}}$ *with* $f_i = \hat{f}_i^{\mathsf{prop}}$ *where*

$$\hat{f}_i^{\mathsf{prop}} = \frac{\boldsymbol{e}_i^T \hat{\mathcal{N}} \boldsymbol{e}_i}{\boldsymbol{e}_i^T \hat{D} \boldsymbol{e}_i},$$

and

$$\begin{aligned} \hat{\mathcal{N}} &= \tau \hat{G}_{1,-1,\beta} \boldsymbol{A} \hat{H}_{-1,1+\beta,\beta} \boldsymbol{A}^T \hat{G}_{1,-1,\beta} \\ \hat{D} &= \hat{\mathcal{N}} + j(\hat{\ell}_0^{\mathsf{e}}) \sum_{j=1}^{K} \hat{\zeta}_j^{-4} \Big(h_{1,-1,\beta}(\hat{\ell}_j^{\mathsf{e}}) - h_{1,-1,\beta}(\hat{\ell}_0^{\mathsf{e}}) \Big)^2 \boldsymbol{A} \hat{\boldsymbol{p}}_j (\boldsymbol{A} \hat{\boldsymbol{p}}_j)^T, \end{aligned}$$

with $j(x) \coloneqq x + \tau x^\beta$ *as the scalar version of* $\mathcal{J}(\boldsymbol{\Sigma})$.

The substitution principle discussed above is based on consistently measuring the principal eigen values and eigen vectors of $\boldsymbol{\Sigma}$ in (7) using equations (9) to (12). Next, we establish the analogue of Theorem 1 for this setup. The proof is much more complicated, as for a general $\boldsymbol{A}$, the expression in the posterior covariances loses commutativity in multiplicative operations between $\boldsymbol{A}$ and $\boldsymbol{\Sigma}$. The result is that for quadratic form estimation, we need to be precise in tackling the distortion in the spectrum of the posterior variance due to the presence of the linear aggregation matrix $\boldsymbol{A}$. We show that the substitution principle, which avoids higher order corrections, is asymptotically consistent, and as good as in the setting of known $\boldsymbol{\Sigma}$, when $p = O(\sqrt{n})$, which is a scenario encountered in many practical applications. However, if $n^{-1/2} p \to \infty$, the approximation error for the substitution rule is suboptimal due to the cost of inversion paid by the simple substitution rule.

Theorem 3 (Asymptotic consistency of $\boldsymbol{b}^T \hat{G}_{r,\alpha,\beta} \boldsymbol{b}$)**.** *Under assumptions* ***A1****,* ***A2****,* ***A3*** *and* ***A4****, uniformly over* $\tau \in \boldsymbol{T}_0$*,* $\beta \in \boldsymbol{B}_0$ *and* $\boldsymbol{b} \in \mathcal{B}$ *with* $||\boldsymbol{b}||_2 = 1$*, we have for all* $(r, \alpha) \in \{-1, 0, 1\} \times \mathbb{R}$

$$\sup_{\tau \in \boldsymbol{T}_0, \beta \in \boldsymbol{B}_0, \boldsymbol{b} \in \mathcal{B}} \Big| \boldsymbol{b}^T \hat{G}_{r,\alpha,\beta} \boldsymbol{b} - \boldsymbol{b}^T G_{r,\alpha,\beta} \boldsymbol{b} \Big| = O_p \Big\{ \max \Big(\frac{p}{n}, \sqrt{\frac{\log n}{n}} \Big) \Big\},$$

where the dependence of $G_{r,\alpha,\beta}$ *on* τ *has been kept implicit for notational ease.*

Theorem 4 (Oracle optimality of CASP)**.** *Under assumptions* ***A1****,* ***A2****,* ***A3*** *and* ***A4****, and the hierarchical model of equations* (1) *and* (2)*, we have, conditionally on* $\boldsymbol{X}$*,*

$$\sup_{\tau \in \boldsymbol{T}_0, \beta \in \boldsymbol{B}_0} \frac{\Big\| \hat{\boldsymbol{q}}^{\mathsf{casp}}(\boldsymbol{AX} | \boldsymbol{S}, \hat{\boldsymbol{f}}^{\mathsf{prop}}, \tau, \beta) - \hat{\boldsymbol{q}}^{\mathsf{cs}}(\boldsymbol{AX} | \boldsymbol{S}, \boldsymbol{f}^{\mathsf{OR}}, \tau, \beta) \Big\|_2^2}{\Big\| \hat{\boldsymbol{q}}^{\mathsf{cs}}(\boldsymbol{AX} | \boldsymbol{S}, \boldsymbol{f}^{\mathsf{OR}}, \tau, \beta) - \boldsymbol{e}_i^T \boldsymbol{A} \boldsymbol{\eta}_0 \Big\|_2^2} = O_p \Big\{ \max \Big(\Big(\frac{p}{n} \Big)^2, \frac{\log n}{n} \Big) \Big\}.$$

Using Theorem 4, we show that in the aggregated model too the data driven adaptive shrinkage factors $\hat{f}_i^{\mathsf{prop}}$ continue to guarantee the oracle optimality of $\hat{\boldsymbol{q}}^{\mathsf{casp}}$ in the class $\mathcal{Q}^{\mathsf{cs}}_{\boldsymbol{A}}$. For the proofs of Theorem 3 and 4, we refer the reader to appendices A.4 and A.8, respectively.

Remark 1. On the uncertainty in estimating K - For estimating the number of spikes in equation (7), we use the procedure in Kritchman and Nadler (2009). Theorem 5 of Kritchman and Nadler (2009) shows that if the level parameter $\alpha_m \downarrow 0$ as $m \uparrow \infty$ such that the $(1-\alpha_m)$th quantile of the Tracy-Widom distribution diverges at a rate slower than $\sqrt{m}$, then $P(\hat{K} = K) \to 0$ as $m \to \infty$. By this result, Theorems 1, 3, 2 and 4 which involve convergence in probability, easily extend to scenarios where K is unknown and is estimated from the data using α_m as prescribed in Theorem 5 of Kritchman and Nadler (2009).

Remark 2. Implementation and R package `casp` - The R package `casp` has been developed to implement our proposed CASP methodology in aggregated as well as disaggregated prediction problems. It is publicly available at the following GitHub repository: `https://github.com/trambakbanerjee/casp`. For estimating the prior hyper-parameters (τ, β) and the number of spikes K, the approach described in Section 3.3 is followed.

5. Simulation studies

In this section we asses the predictive performance of CASP across a wide range of simulation experiments. We consider four competing predictive rules that use different methodologies to estimate $\boldsymbol{\Sigma}$ and thereafter plug-in their respective estimates of $\boldsymbol{\Sigma}$ in the Bayes predictive rule of Lemma 1. In what follows, we briefly discuss these competing methods for estimating $\boldsymbol{\Sigma}$:

1. $\hat{\boldsymbol{q}}^{\mathsf{Bcv}}$ - the predictive rule that uses the bi-cross-validation approach of Owen and Wang (2016) which, under a heteroscedastic factor model structure, first estimates the number of factors, then constructs an estimate $\boldsymbol{S}^{\mathsf{Bcv}}$ of $\boldsymbol{\Sigma}$ and finally plugs-in $\boldsymbol{S}^{\mathsf{Bcv}}$ in the Bayes predictive rule of Lemma 1. We use the implementation available in the R package `esaBcv` for our simulations.

2. $\hat{\boldsymbol{q}}^{\mathsf{Fact}}$ - the predictive rule that uses the FactMLE algorithm of Khamaru and Mazumder (2019) to obtain an estimate $\boldsymbol{S}^{\mathsf{Fact}}$ of $\boldsymbol{\Sigma}$ by formulating the low rank maximum likelihood Factor Analysis problem as a non-linear, non-smooth semidefinite optimization problem. The implementation of the FactMLE algorithm is available in the R package `FACTMLE` wherein we use an estimate $\hat{K}$ of K as discussed in Section 4.

3. $\hat{\boldsymbol{q}}^{\mathsf{Poet}}$ - the predictive rule that uses the approach of Fan et al. (2013) to obtain $\boldsymbol{S}^{\mathsf{Poet}}$ by first retaining the first $\hat{K}$ principal components of $\boldsymbol{S}$ and then uses a thresholding procedure on the remaining sample covariance matrix $\boldsymbol{S}$. The implementation of this approach is available in the R-package `POET` where $\hat{K}$ is an estimate of the number of spikes from Section 4.

4. $\hat{\boldsymbol{q}}^{\mathsf{Naive}}$ - the Naive predictive rule which first estimates the number of spikes $\hat{K}$ from the data, reconstructs the sample covariance matrix $\boldsymbol{S}^{\mathsf{Naive}}$ from the leading $\hat{K}$ eigenvalues and eigenvectors of $\boldsymbol{S}$, and finally plugs in $\boldsymbol{S}^{\mathsf{Naive}}$ in place of $\boldsymbol{\Sigma}$ in the Bayes predictive rule $\boldsymbol{q}^{\mathsf{Bayes}}$ in Lemma 1.

To assess the performance of various predictive rules, we calculate a relative estimation error (REE) which is defined as

$$\text{REE}(\hat{\boldsymbol{q}}) = \frac{R_p(\boldsymbol{\psi}, \hat{\boldsymbol{q}}) - R_p(\boldsymbol{\psi}, \boldsymbol{q}^{\mathsf{Bayes}})}{R_p(\boldsymbol{\psi}, \hat{\boldsymbol{q}}^{\mathsf{S}}) - R_p(\boldsymbol{\psi}, \boldsymbol{q}^{\mathsf{Bayes}})},$$

where for any prediction rule $\hat{\boldsymbol{q}}$ of $\boldsymbol{\psi} = \boldsymbol{A}\boldsymbol{\theta}$, $R_p(\boldsymbol{\psi}, \hat{\boldsymbol{q}})$ is the predictive risk of $\hat{\boldsymbol{q}}$ as discussed in Section 2.3, $\hat{\boldsymbol{q}}^{\mathsf{S}}$ is CASP with shrinkage factors $f_i = 1$ for all i and $\boldsymbol{q}^{\mathsf{Bayes}}$ is the Bayes predictive rule based on the knowledge of unknown $\boldsymbol{\Sigma}$. A value of REE larger than 1 implies poorer prediction performance of $\hat{\boldsymbol{q}}$ relative to $\hat{\boldsymbol{q}}^{\mathsf{S}}$ whereas a value smaller than 1 implies a better prediction performance. In particular, REE allows us to quantify the relative advantage of using coordinate wise adaptive shrinkage in our proposed predictive rule $\hat{\boldsymbol{q}}^{\mathsf{casp}}$.

The R scripts that reproduce the results of our numerical experiments in Sections 5.1 to 5.3 are available at `https://github.com/trambakbanerjee/CASP_paper`.

5.1 Experiment 1

In the setup of experiment 1 we investigate the prediction performance of the five predictive rules under the disaggregated model ($\boldsymbol{A} = \boldsymbol{I}_n$) and sample $\boldsymbol{\theta}$ from an $n = 200$ variate Gaussian distribution with mean vector $\boldsymbol{\eta}_0 = \boldsymbol{0}$ and covariance $\tau\boldsymbol{\Sigma}^{\beta}$. We impose a spike covariance structure on $\boldsymbol{\Sigma}$ with $K = 10$ spikes under the following two scenarios with l_0 fixed at 1.

- Scenario 1: we consider the generalized absolute loss function in equation (3) with b_i sampled uniformly between $(0.9, 0.95)$, $h_i = 1 - b_i$ with $(\tau, \beta) = (0.5, 0.25)$ and K spikes equi-spaced between 80 and 20.
- Scenario 2: we consider the Linex loss function in equation (4) with a_i sampled uniformly between $(-2, -1)$, $b_i = 1$ with $(\tau, \beta) = (1, 1.75)$ and K spikes equi-spaced between 25 and 5.

To estimate $\boldsymbol{S}$, we sample $\boldsymbol{W}_j$ independently from $N_n(\boldsymbol{0}, \boldsymbol{\Sigma})$ for $j = 1, \ldots, m$, where we allow m to vary over $(15, 20, 25, 30, 35, 40, 45, 50)$. Finally $m_x = 1$ copy of $\boldsymbol{X}$ is sampled from $N_n(\boldsymbol{\theta}, \boldsymbol{\Sigma})$ with $m_0 = 1$. This sampling scheme is repeated over 500 repetitions and the REE of the competing predictive rules and CASP is presented in figures 1 and 2 for scenarios 1 and 2, respectively. In table 1, we report the REE and the estimates of (K, τ, β) at $m = 15$. Using the R-package `POET`, the estimation of $\boldsymbol{S}^{\mathsf{Poet}}$ was extremely slow in our simulations and therefore we report the REE of $\hat{\boldsymbol{q}}^{\mathsf{Poet}}$ only at $m = 15$ and exclude this predictive rule from the figures. The left panels of figures 1 and 2 both suggest a superior risk performance of CASP as m varies. Moreover, when the ratio $n/(m-1)$ is largest, the right panels of these figures plot the sorted shrinkage factors $\hat{f}_i^{\mathsf{prop}}$ averaged over the 500 repetitions (red line) and sandwiched between its 10^{th} and 90^{th} percentiles (represented by the gray shaded region) under the two scenarios. Under scenario 1 in particular, the estimated shrinkage factors are all smaller than 1 indicating the significant role that the coordinate-wise shrinkage plays in reducing the marginal mean square error of $\hat{\boldsymbol{q}}^{\mathsf{casp}}$ from $\boldsymbol{q}^{\mathsf{Bayes}}$. However as β increases from 0.25 to 1.75 in scenario 2, the estimated shrinkage factors move closer to 1, and the risk performances of $\hat{\boldsymbol{q}}^{\mathsf{casp}}$ and $\hat{\boldsymbol{q}}^{\mathsf{S}}$ are indistinguishable from each other as seen in table 1

Table 1: Relative Error estimates (REE) of the competing predictive rules at $m = 15$ for Scenarios 1 and 2 under Experiment 1. The numbers in parenthesis are standard errors over 500 repetitions.

	Scenario 1: $(K, \tau, \beta) = (10, 0.5, 0.25)$				Scenario 2: $(K, \tau, \beta) = (10, 1, 1.75)$			
	$\hat{K}$	$\hat{\tau}$	$\hat{\beta}$	REE($\hat{\boldsymbol{q}}$)	$\hat{K}$	$\hat{\tau}$	$\hat{\beta}$	REE($\hat{\boldsymbol{q}}$)
CASP	7 (0.04)	0.59 (0.002)	0.27 (0.002)	**0.95**	4 (0.04)	0.97 (0.004)	1.79 (0.006)	**1.00**
Bcv	3 (0.08)	0.58 (0.003)	0.26 (0.003)	1.14	1 (0.04)	1.00 (0.001)	1.75 (0.006)	4.24
FactMLE	7 (0.04)	0.57 ($< 10^{-3}$)	0.19 (0.001)	1.68	4 (0.04)	0.98 ($< 10^{-3}$)	1.55 (0.001)	4.58
POET	7 (0.04)	0.57 ($< 10^{-3}$)	0.18 ($< 10^{-3}$)	2.14	4 (0.04)	0.97 ($< 10^{-3}$)	1.53 ($< 10^{-3}$)	7.26
Naive	7 (0.04)	0.60 ($< 10^{-3}$)	0.24 (0.001)	1.36	4 (0.04)	1.00 ($< 10^{-3}$)	1.63 (0.004)	1.87

under scenario 2 wherein the REE of CASP is 1. This is not unexpected because with a fixed $\tau > 0$ and β growing above 1, the factor $\sum_{j=1}^{K} \hat{\varsigma}_j^{-4} \Big(h_{1,-1,\beta}(\hat{\ell}_j^{\mathsf{e}}) - h_{1,-1,\beta}(\hat{\ell}_0^{\mathsf{e}}) \Big)^2$ in the denominator of $\hat{f}_i^{\mathsf{prop}}$ becomes smaller in comparison to the numerator $\hat{\mathcal{N}}$ in Definition 4 and the improvement due to coordinate-wise shrinkage dissipates. From table 1, we see

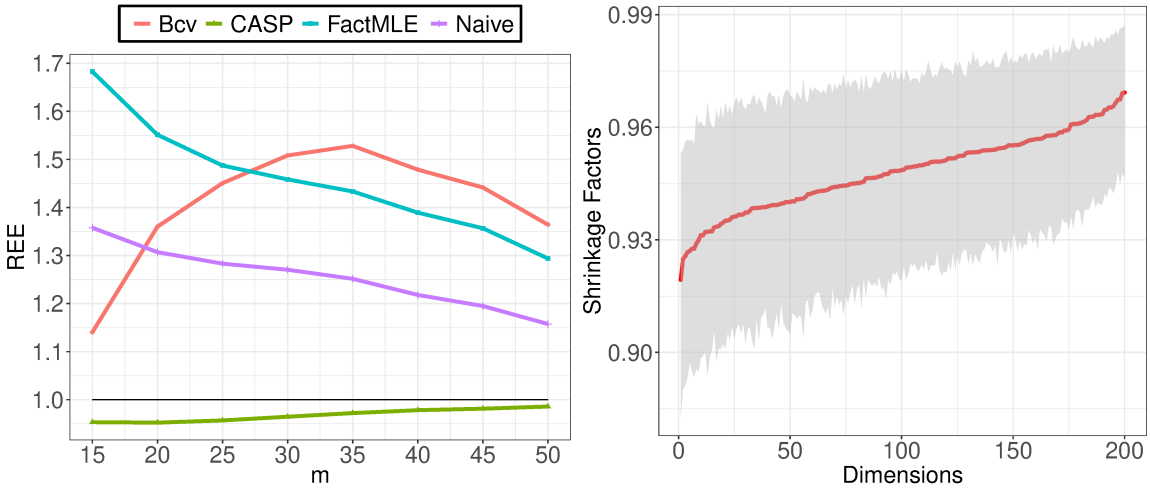

Figure 1: Experiment 1 Scenario 1 (Generalized absolute loss): Left - Relative Error estimates as m varies over $(15, 20, 25, 30, 35, 40, 45, 50)$. Right: Magnitude of the sorted shrinkage factors $\hat{f}_i^{\mathsf{prop}}$ averaged over 500 repetitions at $m = 15$ and sandwiched between its 10^{th} and 90^{th} percentiles

that $\hat{\boldsymbol{q}}^{\mathsf{Bcv}}$ is the most competitive predictive rule next to $\hat{\boldsymbol{q}}^{\mathsf{casp}}$ across both the scenarios however it seems to suffer from the issue of under estimation of the number of factors K. We notice this behavior of $\hat{\boldsymbol{q}}^{\mathsf{Bcv}}$ across all our numerical and real data examples.

The other three predictive rules, $\hat{\boldsymbol{q}}^{\mathsf{Fact}}$, $\hat{\boldsymbol{q}}^{\mathsf{Poet}}$ and $\hat{\boldsymbol{q}}^{\mathsf{Naive}}$, exhibit poorer risk performances and this is not entirely surprising in this setting primarily because the four competing predictive rules considered here do not involve any asymptotic corrections to the sample eigenvalues and their eigenvectors whereas CASP uses the phase transition phenomenon of the sample eigenvalues and their eigenvectors to constructs consistent estimators of smooth functions of $\boldsymbol{\Sigma}$ that appear in the form of the Bayes predictive rules.

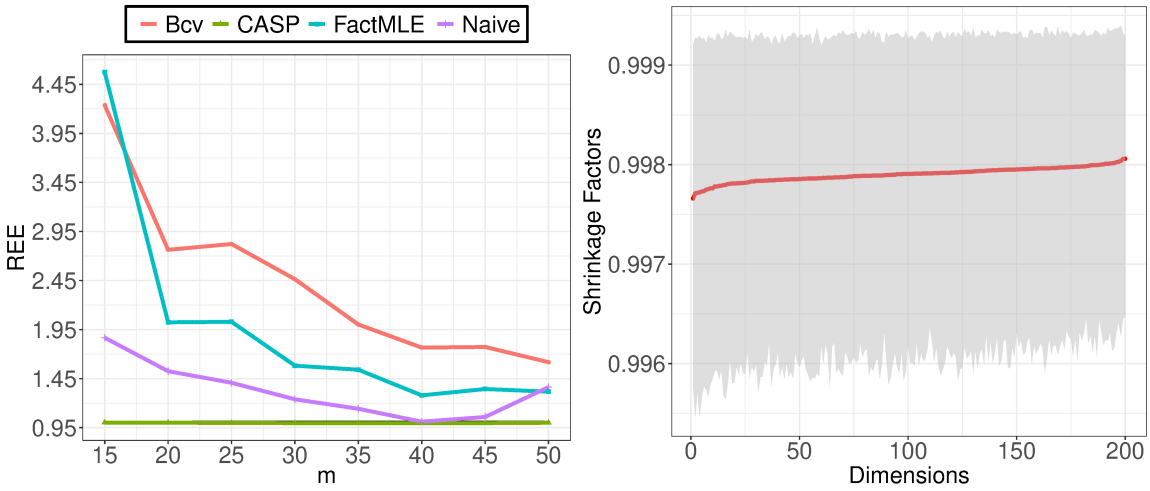

Figure 2: Experiment 1 Scenario 2 (Linex loss): Left - Relative Error estimates as m varies over $(15, 20, 25, 30, 35, 40, 45, 50)$. Right: Magnitude of the sorted shrinkage factors $\hat{f}_i^{\text{prop}}$ averaged over 500 repetitions at $m = 15$ and sandwiched between its 10^{th} and 90^{th} percentiles

From figure 1, we see that as m increases, $\hat{\boldsymbol{q}}^{\text{Naive}}$ performs better than $\hat{\boldsymbol{q}}^{\text{Bcv}}$. The spiked covariance structure considered in scenario 1 is substantially strong as there are $K = 10$ equispaced spikes between 20 and 80. The $\hat{\boldsymbol{q}}^{\text{Bcv}}$ method, which underestimates K more severely compared to $\hat{\boldsymbol{q}}^{\text{Naive}}$, performs worse as m increases for there is more information to estimate the spiked structure. The same phenomenon happens in scenario 2 where $\beta > 1$. However, when $\beta > 1$, most of the coordinate-wise shrinkage factors are close to 1 (see the right plot of figure 2) and so, the difference between CASP and $\hat{\boldsymbol{q}}^{\text{Naive}}$ is not much due to coordinate-wise shrinkage but mostly due to the biased estimation of the eigenvalues by the naive method.

5.2 Experiment 2

For experiment 2 we consider the setup of a static factor model with heteroscedastic noise and simulate our data according to the following model:

$$\begin{aligned} \boldsymbol{X}_t &= \boldsymbol{\theta} + \boldsymbol{B}\boldsymbol{\Gamma}_t + \boldsymbol{\epsilon}_t \\ \boldsymbol{\Gamma}_t &\sim N_K(\boldsymbol{0}, \boldsymbol{I}_K) \\ \boldsymbol{\theta} &\sim N_n(\boldsymbol{\eta}_0, \tau\boldsymbol{\Sigma}^\beta) \text{ and } \boldsymbol{\epsilon}_t \sim N_n(\boldsymbol{0}, \boldsymbol{\Delta}_n), \end{aligned}$$

where $K \ll n$ represents the number of latent factors, $\boldsymbol{B}$ is the $n \times K$ matrix of factor loadings, $\boldsymbol{\Gamma}_t$ is the $K \times 1$ vector of latent factors independent of $\boldsymbol{\epsilon}_t$ and $\boldsymbol{\Delta}_n$ is an $n \times n$ diagonal matrix of heteroscedastic noise variances. In this model $\boldsymbol{\Sigma} = \boldsymbol{B}\boldsymbol{B}^T + \boldsymbol{\Delta}_n$ and coincides with the heteroscedastic factor models considered in Owen and Wang (2016); Fan et al. (2013); Khamaru and Mazumder (2019) for estimating $\boldsymbol{\Sigma}$. Thus the three competing predictive rules $\hat{\boldsymbol{q}}^{\text{Bcv}}$, $\hat{\boldsymbol{q}}^{\text{Poet}}$ and $\hat{\boldsymbol{q}}^{\text{Fact}}$ are well suited for prediction in this model. Factor models of this form are often considered in portfolio risk estimation (see for example Fan

et al. (2019)) where the goal is to first estimate the matrix of factor loadings $\boldsymbol{B}$ and the vector of latent factors $\boldsymbol{\Gamma}_t$ and thereafter use the fitted model to sequentially predict $\boldsymbol{A}\boldsymbol{X}_{t+s}$ for $s = 1, 2, \ldots, T$ where $\boldsymbol{X}_t$ might represent an n dimensional vector of stock excess returns and $\boldsymbol{A}$ is the $p \times n$ weight matrix that aggregates the predicted excess returns into $p \ll n$ individual portfolios level returns. Often an autoregressive structure is imposed on $\boldsymbol{\Gamma}_t$ so that $\boldsymbol{\Gamma}_t = \boldsymbol{\Phi}\boldsymbol{\Gamma}_{t-1} + \boldsymbol{v}_t$ which is the so called dynamic factor model (Geweke, 1977) where $\boldsymbol{\Phi}$ is a $K \times K$ matrix of autoregressive coefficients and $\boldsymbol{v}_t \sim N_K(\boldsymbol{0}, \boldsymbol{D})$.

For the purposes of this simulation exercise we take $n = 200$, $\boldsymbol{\Phi} = \boldsymbol{0}$ and $\boldsymbol{D} = \boldsymbol{I}_K$ with $K = 10$ factors. We fix $\boldsymbol{\eta}_0 = 0$, and simulate the rows of $\boldsymbol{B}$ independently from $N_K(\boldsymbol{0}, c\boldsymbol{I}_K)$ and the diagonal elements of $\boldsymbol{\Delta}_n$ independently from $\mathsf{Unif}(0.5, 1.5)$. The elements of the aggregation matrix $\boldsymbol{A}$ are simulated uniformly from $(0, 1)$ with $p = 20$ rows normalized to 1. In this experiment, similar to experiment 1, we consider two scenarios:

- Scenario 1: we fix $(c, \tau, \beta) = (0.5, 0.5, 0.25)$.
- Scenario 2: we fix $(c, \tau, \beta) = (0.2, 1.5, 2)$.

To estimate $\boldsymbol{S}$, we sample $\boldsymbol{W}_j$ independently from $N_n(\boldsymbol{0}, \boldsymbol{\Sigma})$ for $j = 1, \ldots, m$, where we allow m to vary over $(15, 20, 25, 30, 35, 40, 45, 50)$. Finally $m_x = 1$ copy of $\boldsymbol{X}_t$ is sampled from $N_n(\boldsymbol{\theta}, \boldsymbol{\Sigma})$ and the goal is to predict $\boldsymbol{A}\boldsymbol{X}_{t+1}$ under a Linex loss with $b_i = 1$ and $a_i \overset{i.i.d}{\sim} \mathsf{Unif}(1, 2)$ to emphasize the severity of overprediction of portfolio excess returns. This sampling scheme is repeated over 500 repetitions and the REE of the competing predictive rules and CASP is presented in figures 3 and 4 for scenarios 1 and 2, respectively. From

Table 2: Relative Error estimates (REE) of the competing predictive rules at $m = 15$ for Scenarios 1 and 2 under Experiment 2. The numbers in parenthesis are standard errors over 500 repetitions.

	Scenario 1: $(K, \tau, \beta) = (10, 0.5, 0.25)$				Scenario 2: $(K, \tau, \beta) = (10, 1.5, 2)$			
	$\hat{K}$	$\hat{\tau}$	$\hat{\beta}$	REE($\hat{\boldsymbol{q}}$)	$\hat{K}$	$\hat{\tau}$	$\hat{\beta}$	REE($\hat{\boldsymbol{q}}$)
CASP	8 (0.04)	0.59 (0.002)	0.28 (0.002)	0.91	2 (0.04)	1.48 (0.004)	2.09 (0.003)	**1.00**
Bcv	3 (0.08)	0.58 (0.003)	0.26 (0.002)	**0.87**	0 (0.01)	1.50 ($< 10^{-3}$)	2.04 (0.004)	1.23
FactMLE	8 (0.04)	0.57 ($< 10^{-3}$)	0.19 (0.001)	1.49	2 (0.04)	1.48 ($< 10^{-3}$)	1.87 (0.002)	1.10
POET	8 (0.04)	0.57 ($< 10^{-3}$)	0.18 ($< 10^{-3}$)	1.87	2 (0.04)	1.47 ($< 10^{-3}$)	1.83 ($< 10^{-3}$)	1.12
Naive	8 (0.04)	0.60 ($< 10^{-3}$)	0.24 (0.001)	1.61	2 (0.04)	1.50 ($< 10^{-3}$)	2.05 (0.004)	1.27

the left plot in figure 3 we see that $\hat{\boldsymbol{q}}^{\mathsf{Bcv}}$ returns the smallest REE amongst all competing predictive rules in scenario 1 and is closely followed by $\hat{\boldsymbol{q}}^{\mathsf{casp}}$. This is expected since $\hat{\boldsymbol{q}}^{\mathsf{Bcv}}$ relies on a heteroscedastic factor model structure to estimate $\boldsymbol{\Sigma}$, however, even in this scenario CASP is competitive. In particular, the estimated shrinkage factors for CASP are all smaller than 1 (right plot in figure 3) which allows CASP to deliver an REE which is substantially less than 1. Under scenario 2 (left plot in figure 4) $\hat{\boldsymbol{q}}^{\mathsf{Bcv}}$ no longer enjoys a superior performance and exhibits a volatile REE profile as m increases from 15 to 50 which can potentially be due to its tendency to under estimate the number of factors K as seen from table 2 at $m = 15$. In this scenario, the $\hat{\boldsymbol{q}}^{\mathsf{casp}}$ is the most competitive predictive rule however with $\beta > 1$, $\hat{\boldsymbol{q}}^{\mathsf{casp}}$ is no better than $\hat{\boldsymbol{q}}^{\mathsf{S}}$ in terms of REE.

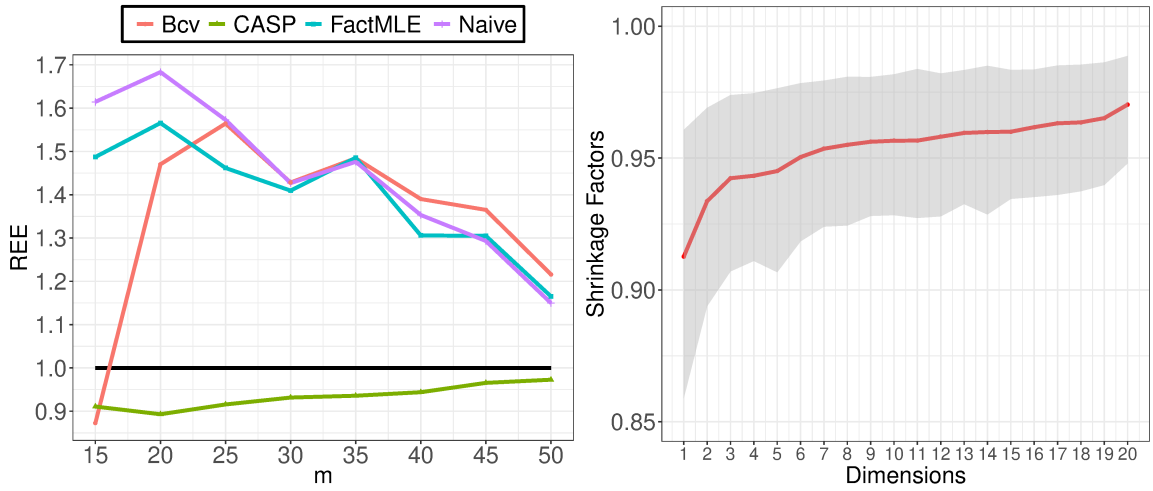

Figure 3: Experiment 2 Scenario 1 (Linex loss): Left - Relative Error estimates as m varies over $(15, 20, 25, 30, 35, 40, 45, 50)$. Right: Magnitude of the sorted shrinkage factors $\hat{f}_i^{\text{prop}}$ averaged over 500 repetitions at $m = 15$ and sandwiched between its 10^{th} and 90^{th} percentiles

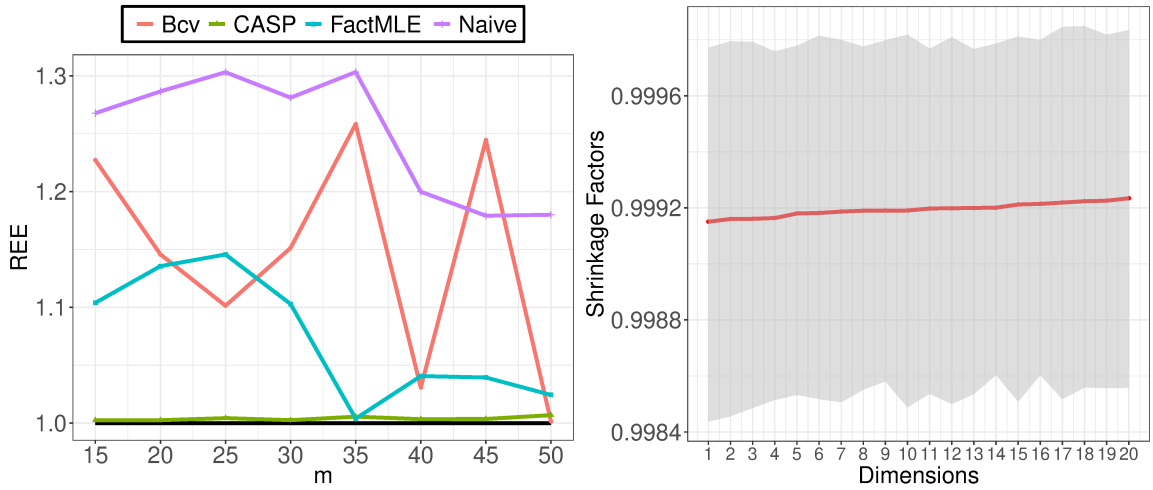

Figure 4: Experiment 2 Scenario 2 (Linex loss): Left - Relative Error estimates as m varies over $(15, 20, 25, 30, 35, 40, 45, 50)$. Right: Magnitude of the sorted shrinkage factors $\hat{f}_i^{\text{prop}}$ averaged over 500 repetitions at $m = 15$ and sandwiched between its 10^{th} and 90^{th} percentiles

5.3 Experiment 3

For experiment 3, we consider a slightly different setup where we do not impose a spike covariance structure on $\boldsymbol{\Sigma}$. Instead, we assume that $(\boldsymbol{\Sigma})_{ij} = Cov(X_i, X_j) = 0.9^{|i-j|}$ where $i, j = 1, \ldots, n$, thus imposing an AR(1) structure between the n coordinates of $\boldsymbol{X}$. As in experiment 1, we sample $\boldsymbol{\theta}$ from an $n = 200$ variate Gaussian distribution with mean vector

$\boldsymbol{\eta}_0 = \mathbf{0}$ and covariance $\tau\boldsymbol{\Sigma}^\beta$. We vary (τ, β) across two scenarios where we take (τ, β) as $(1, 0.5)$ and $(0.5, 2)$ in scenarios 1 and 2 respectively. We estimate $\boldsymbol{S}$ using the approach described in experiments 1 and 2, and sample $m_x = 1$ copy of $\boldsymbol{X}$ from $N_n(\boldsymbol{\theta}, \boldsymbol{\Sigma})$ with a goal to predict $\boldsymbol{AY}$ under a generalized absolute loss function with $h_i = 1 - b_i$ and b_i sampled uniformly from $(0.9, 0.95)$ for $i = 1, \cdots, p$. Here $\boldsymbol{Y} \sim N_n(\boldsymbol{\theta}, \boldsymbol{\Sigma})$ is independent of $\boldsymbol{X}$ and $\boldsymbol{A}$ is a fixed $p \times n$ sparse matrix with the $p = 20$ rows sampled independently from a mixture distribution with density $0.9\delta_0 + 0.1\,Unif(0, 1)$ and normalized to 1 thereafter. This sampling scheme is repeated over 500 repetitions and the REE of the competing predictive rules and CASP is presented in figures 5, 6 and table 3.

Table 3: Relative Error estimates (REE) of the competing predictive rules at $m = 15$ for Scenarios 1 and 2 under Experiment 3. The numbers in parenthesis are standard errors over 500 repetitions.

	Scenario 1: $(\tau, \beta) = (1, 0.5)$				Scenario 2: $(\tau, \beta) = (0.5, 2)$			
	$\hat{K}$	$\hat{\tau}$	$\hat{\beta}$	REE($\hat{\boldsymbol{q}}$)	$\hat{K}$	$\hat{\tau}$	$\hat{\beta}$	REE($\hat{\boldsymbol{q}}$)
CASP	7 (0.06)	1.09 (0.007)	0.40 (0.004)	**0.94**	7 (0.06)	0.57 ($< 10^{-3}$)	1.74 (0.001)	**1.00**
Bcv	1 (0.08)	0.95 (0.012)	0.36 (0.002)	2.39	1 (0.08)	0.59 ($< 10^{-3}$)	1.79 (0.002)	4.27
FactMLE	7 (0.06)	1.16 (0.001)	0.33 ($< 10^{-3}$)	1.23	7 (0.06)	0.57 ($< 10^{-3}$)	1.73 ($< 10^{-3}$)	1.08
POET	7 (0.06)	1.17 ($< 10^{-3}$)	0.33 ($< 10^{-3}$)	1.49	7 (0.06)	0.57 ($< 10^{-3}$)	1.73 ($< 10^{-3}$)	1.27
Naive	7 (0.06)	1.16 (0.001)	0.34 (0.001)	1.26	7 (0.06)	0.57 ($< 10^{-3}$)	1.73 ($< 10^{-3}$)	1.11

In this setup, the departure from the factor model leads to a poorer estimate of β for CASP than what was observed under experiments 1 and 2, however, the REE of CASP continues to be the smallest amongst all the other competing rules. When $\beta = 2$ (scenario 2), $\hat{\boldsymbol{q}}^{\mathsf{casp}}$ and $\hat{\boldsymbol{q}}^{\mathsf{S}}$ are almost identical in their performance. Amongst the competing methods

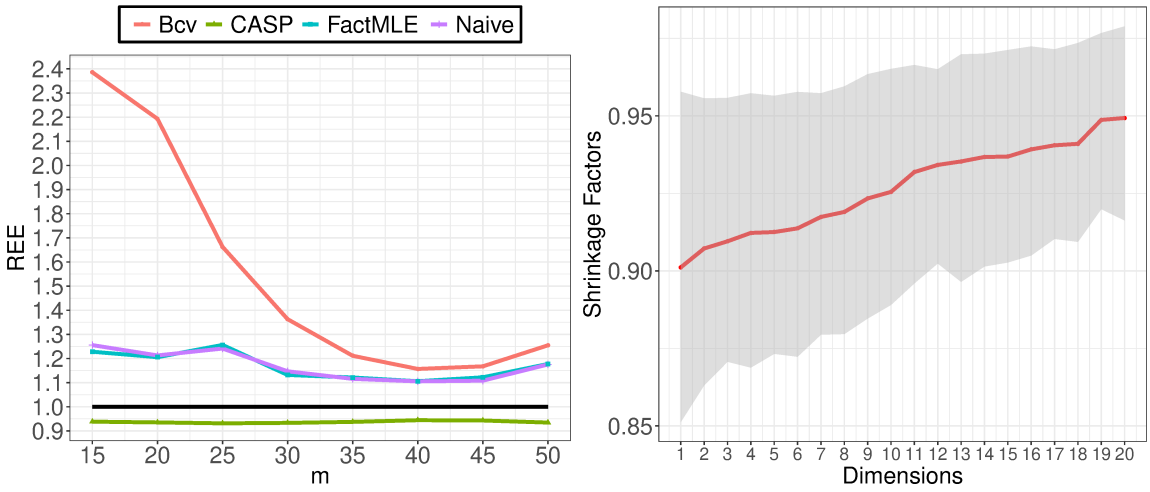

Figure 5: Experiment 3 Scenario 1 (Generalized absolute loss): Left - Relative Error estimates as m varies over $(15, 20, 25, 30, 35, 40, 45, 50)$. Right: Magnitude of the sorted shrinkage factors $\hat{f}_i^{\mathsf{prop}}$ averaged over 500 repetitions at $m = 15$ and sandwiched between its 10^{th} and 90^{th} percentiles

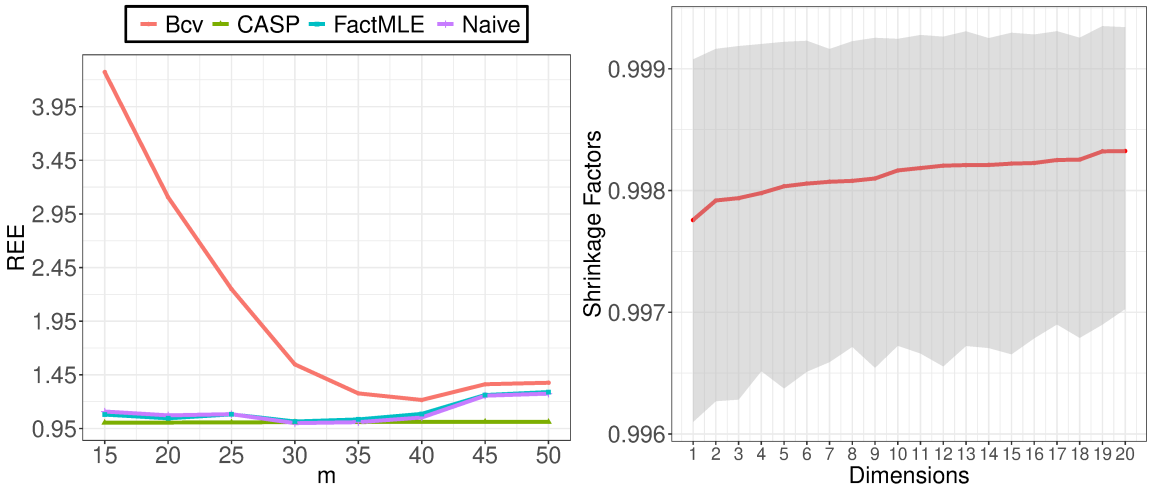

Figure 6: Experiment 3 Scenario 2 (Generalized absolute loss): Left - Relative Error estimates as m varies over $(15, 20, 25, 30, 35, 40, 45, 50)$. Right: Magnitude of the sorted shrinkage factors $\hat{f}_i^{\text{prop}}$ averaged over 500 repetitions at $m = 15$ and sandwiched between its 10^{th} and 90^{th} percentiles

$\hat{\boldsymbol{q}}^{\text{Bcv}}$ has the highest REE, possibly exacerbated by the departure from a factor based model considered in this experiment whereas this seems to have a comparatively lesser impact on CASP indicating potential robustness of CASP to misspecifications of the factor model.

6. Real data illustration with groceries sales data

In this section we analyze a part of the dataset published by Bronnenberg et al. (2008). This dataset has been used in significant studies related to consumer behavior, spending and their policy implications (see for example Bronnenberg et al. (2012); Coibion et al. (2015)). The dataset holds the weekly sales and scanner prices of common grocery items sold in retail outlets across 50 states in the U.S. The retail outlets available in the dataset have identifiers that link them to the city that they serve. In accordance to our lagged data example, we analyze a part of this dataset that spans $m = 100$ weeks from December 31, 2007 to November 29, 2009 as substantial amount of disaggregate data from distant past that will be used for constructing auxiliary information on the covariance. We use 3 weeks from a relatively recent snapshot covering October 31, 2011 to November 20, 2011 as data from the current model. We assume, as in equation (6), that there might have been drift change in the sales data across time but the covariances across stores are invariant over time. Our goal is to predict the state level total weekly sales across all retail outlets for four common grocery items: coffee, mayo, frozen pizza and carbonated beverages. We use the most recent $T = 2$ weeks, from November 7, 2011 to November 20, 2011 as our prediction period and utilize the sales data of week $t - 1$ to predict the state aggregated totals for week t where $t = 1, \ldots, T$. For each of the four products, the prediction period includes sales across approximately $n = 1,140$ retail outlets that vary significantly in terms of their size and quantity sold across the T weeks. Moreover, some of the outlets have undergone

merger and even closure during the prediction period which is often recorded as 0 product sales. Let $\mathbb{X}_0^{(p)}$ be the n dimensional vector denoting the number of units of product p sold across the n outlets in week 0 - October 31, 2011 to November 6, 2011. For our prediction problem, we use a threshold of s_p units for product p and consider only those outlets that have sold at least s_p units in week 0.

Table 4: Loss ratios (16) across six predictive rules for four products.

Product	Method	K	(*a.*) Total sales by state Week 1 Loss Ratio	(*b.*) Total sales by state Week 2 Loss Ratio
Coffee $s_p = 1000,\ n_p = 233,\ d_p = 31$	CASP	26	**0.999**	**1.002**
	Naive	26	1.044	1.063
	Bcv	17	1.043	1.036
	POET	26	1.047	1.070
	FactMLE	26	1.009	1.044
	Unshrunk	-	1.838	2.273
Mayo $s_p = 500,\ n_p = 157,\ d_p = 30$	CASP	26	**0.995**	**1.004**
	Naive	26	0.996	1.016
	Bcv	19	1.040	1.019
	POET	26	0.996	1.022
	FactMLE	26	0.999	1.012
	Unshrunk	-	1.084	2.420
Frozen Pizza $s_p = 1000,\ n_p = 359,\ d_p = 33$	CASP	33	**1.000**	**0.998**
	Naive	33	1.177	1.135
	Bcv	19	1.059	1.091
	POET	33	1.033	1.040
	FactMLE	33	1.008	1.020
	Unshrunk	-	4.424	6.701
Carb. Beverages $s_p = 5000,\ n_p = 410,\ d_p = 33$	CASP	37	**1.003**	**0.984**
	Naive	37	1.065	1.033
	Bcv	20	1.073	1.142
	POET	37	1.065	1.038
	FactMLE	37	1.067	1.059
	Unshrunk	-	3.459	8.885

Let $\boldsymbol{X}_{t-1}^{(p)}$ be the $n_p = \sum_{i=1}^{n} \mathbb{I}(\mathbb{X}_{0,i}^{(p)} \geq s_p)$ dimensional vector denoting the number of units of product p sold across n_p stores in week $t-1$. For a distributor, it is economically important to predict the aggregated demand (future sales) for each US state as intrastate transport of inventories, and transfer of business and tax accounts can be easily executed within state boundaries. The time $t-1$ prediction problem then is to predict $\boldsymbol{V}_t^{(p)} = \boldsymbol{A}^{(p)}\boldsymbol{X}_t^{(p)}$ where $\boldsymbol{A}^{(p)}$ is a $d_p \times n_p$ matrix that aggregates product p sales across n_p stores into d_p unique states across the U.S. To evaluate the performance of any predictive rule $\hat{\boldsymbol{q}}_t^{(p)}$, we use the generalized absolute loss function of equation (3) and calculate the time t ratio of total loss for prediction using $\hat{\boldsymbol{q}}_t^{(p)}$ to the total loss for prediction using CASP with all shrinkage factors $f_i = 1$:

$$\mathcal{L}_t\Big(\boldsymbol{q}_t^{(\mathsf{S,p})}, \boldsymbol{q}_t^{(p)})\Big) = \frac{\sum_{i=1}^{d_p}\Big\{b_i\Big(V_{t,i}^{(p)} - \hat{q}_{t,i}^{(p)}\Big)^+ + h_i\Big(\hat{q}_{t,i}^{(p)} - V_{t,i}^{(p)}\Big)^+\Big\}}{\sum_{i=1}^{d_p}\Big\{b_i\Big(V_{t,i}^{(p)} - \hat{q}_{t,i}^{(\mathsf{S,p})}\Big)^+ + h_i\Big(\hat{q}_{t,i}^{(\mathsf{S,p})} - V_{t,i}^{(p)}\Big)^+\Big\}}, \tag{16}$$

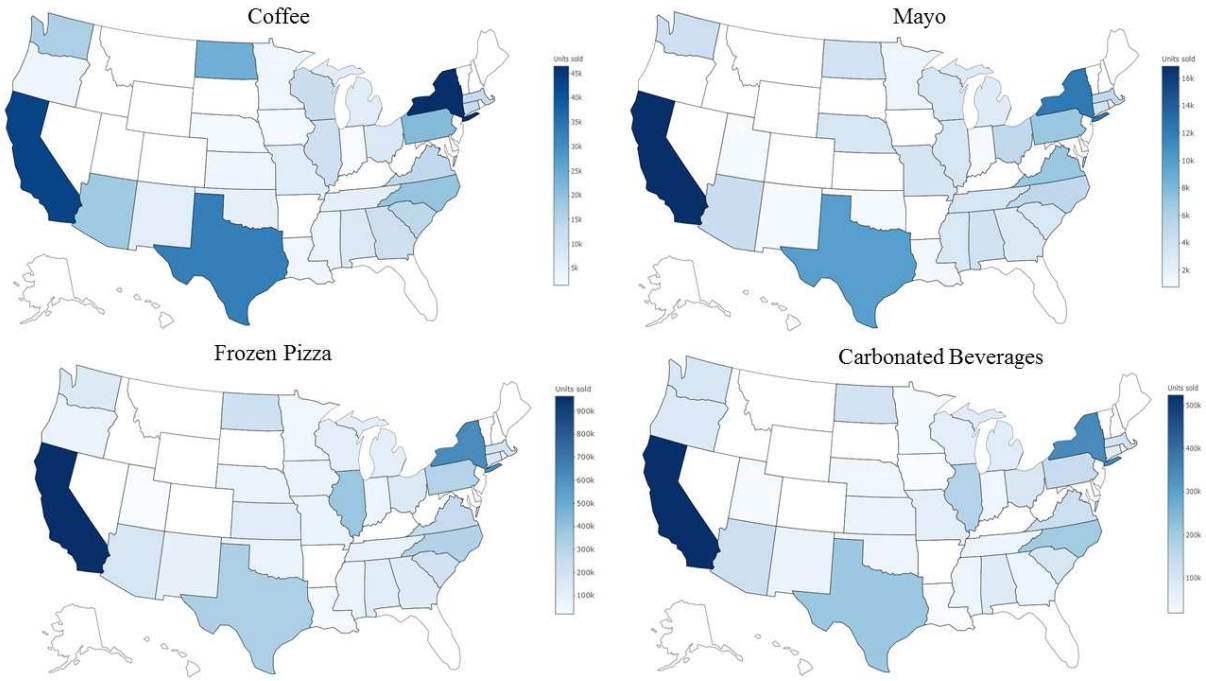

Figure 7: CASP predicted weekly demand of the grocery items across US averaged over the two prediction weeks - November 7, 2011 to November 20, 2011.

where $\hat{\boldsymbol{q}}_t^{(\mathsf{S,p})}$ is CASP with $f_i = 1$ for all i. We set $b_i = 0.95$ and $h_i = 1 - b_i$ for all $i = 1, \ldots, d_p$ to emphasize the severity of underprediction since overstocking may lead to holding and storage costs as all of these four products have considerably longer expiration dates but understocking, on the other hand, may translate into substantial lost sales and reputation costs for the retail outlets. In table 4 we report this loss ratio $\mathcal{L}_t$ for each product p in columns (a) and (b), and for six competing predictive rules: (i) CASP, (ii) the Naive predictive rule as discussed in section 5, (iii) Bcv (Owen and Wang, 2016), (iv) POET (Fan et al., 2013), (v) FactMLE (Khamaru and Mazumder, 2019), and (vi) the Unshrunk predictive rule that simply uses past week's sales to predict the sales in the forthcoming week. To compute an estimate $\boldsymbol{S}^{(p)}$ of the $n_p \times n_p$ population covariance matrix $\boldsymbol{\Sigma}^{(p)}$ of $\boldsymbol{X}_t^{(p)}$ we rely on the additional data on $m = 100$ weeks available from December 31, 2007 to November 29, 2009 and estimate $\boldsymbol{S}^{(p)}$ using the technique described in Section 2.4. In particular, we use the function `smooth.spline` from the R-package `splines2` and choose $k = 3$ knots corresponding to the $25, 50$ and 95 percentiles of the sales distribution across the n_p stores at each of the m weeks. We complete the specification of our model by setting $\boldsymbol{\eta} = \eta_0^{(p)}\mathbf{1}$ where $\eta_0^{(p)}$ is the median of average weekly sales of n_p outlets over the m weeks and use equation (15) to estimate $\beta^{(p)}$ over the interval $[0.1, 1]$ with $\tau^{(p)}$ fixed at 1.

The loss ratios reported in columns (a) and (b) of table 4 indicate a competitive performance of CASP over the five remaining predictive rules. CASP continues to provide the smallest loss ratios across both the weeks with the only exception being the loss ratio in Week 1 (column (a)) for product 'Mayo', where CASP is competitive with the predictive

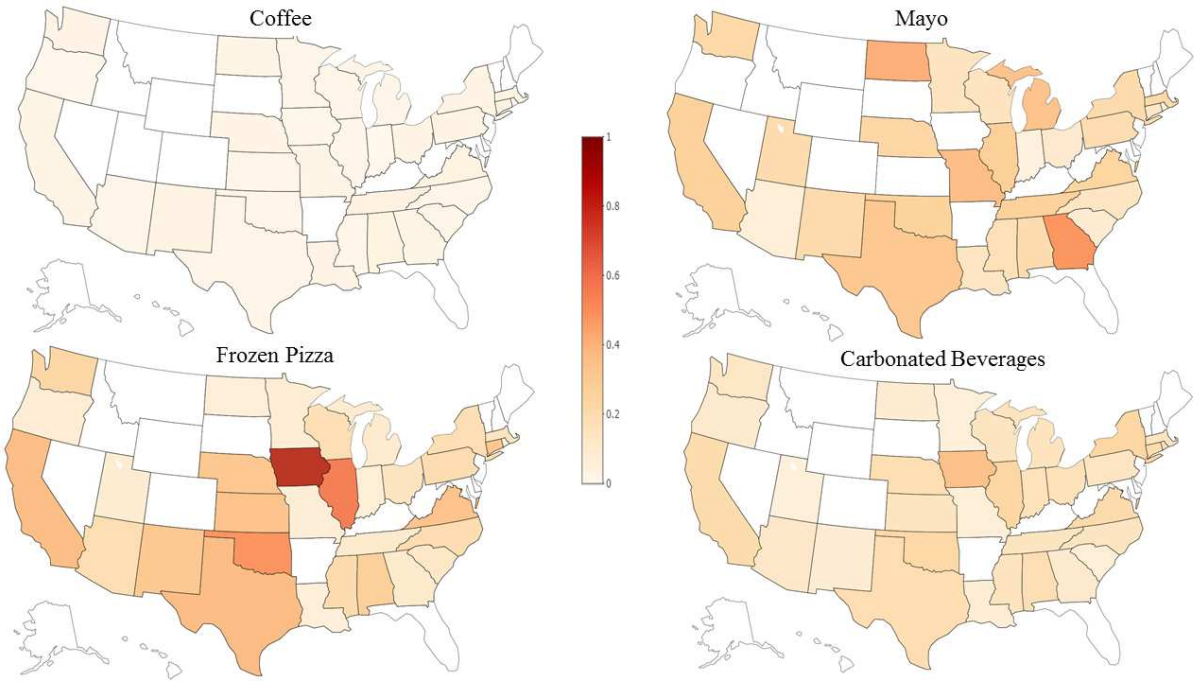

Figure 8: Role of coordinate-wise shrinkage in CASP across US states for four grocery items. In the figures, $1-$ the shrinkage factors is displayed and so, deeper shades denote higher shrinkage.

rules $\hat{\boldsymbol{q}}^{\mathsf{Naive}}$(Naive) and $\hat{\boldsymbol{q}}^{\mathsf{Poet}}$(POET). It is interesting to note that in at least one of the two weeks, the loss ratio of CASP is marginally bigger than 1 across all four product categories, indicating that the coordinate-wise shrinkage factors do not always bring in any significant improvement in prediction performance. This is not entirely unexpected because the hierarchical model assumption of equation (2) may not hold in this setting and thus the model based shrinkage factors $\hat{f}_i^{\mathsf{prop}}$ may not be the most optimal coordinate-wise shrinkage. For each of the four products, figures 7 and 8 present, respectively, the CASP predicted future demand and the distribution of the shrinkage factors of CASP across the different states. In figure 8 the shrinkage factors are plotted as $1 - f_i$ so that a lighter shade in the heatmap corresponds to a smaller shrinkage (larger $\hat{f}_i^{\mathsf{prop}}$). For example in the case of Coffee, the shrinkage factors are all closer to 1 across all the $d_p = 31$ states and this effect translates into loss ratios being almost equal to 1 across the two weeks in table 4. In the case of Frozen Pizza, however, the magnitudes of the shrinkage factors are evenly distributed across the $d_p = 33$ states. For instance, the states of Iowa followed by Illinois and Oklahoma exhibit the largest shrinkages while for Mayo, Georgia and North Dakota have the largest shrinkages in their predicted weekly total sales. In particular, this demonstrates that the shrinkage factors vary across the products because the variability in sales is product specific. More importantly, CASP is flexible enough to capture this inter-product differences and mainly due to its bias correction step (see section 3.1), CASP offers better estimates of future sales than the other popular predictive approaches considered here.

7. Discussion

In this article we considered point prediction in Gaussian models with an unknown covariance that has a spiked structure. Extending our methodology to non-Gaussian models can be useful in a host of applications. Our proposed method depends on eigenvectors associated with the leading eigenvalues of the unknown covariance matrix. In case of standard spiked covariance models, i.e., when $\mathbf{\Sigma}$ is a finite rank perturbation (or a multiple) of identity, recent results of Bloemendal et al. (2016) and Bao et al. (2020) indicate that as long as the first four moments of the observations match with the Gaussian, the behavior of sample spiked eigenvalues, as well as linear functionals involving the leading eigenvectors will be similar, up to the second order, to the Gaussian case. Our proposed predictor depends on the leading eigenvectors only through their linear functions and so, the risk properties of the proposed method stated in Section 3 are likely to hold in such a non-Gaussian setting. However, if the "moment matching condition" does not hold, there may be a need for modifying the proposed predictors based on the moments of the observations. As future work, it will be interesting to comprehensively study the aforementioned prediction problem in the non-Gaussian case. Furthermore, extending our shrinkage methodology beyond the *low rank plus homoscedastic noise* covariance setup of equation (7) to heteroscedastic noise setups will be beneficial for several practical applications. As future work, it will be useful to introspect efficient shrinkage predictors under such covariance structures.

Acknowledgments

We thank the Action Editor and three anonymous referees whose comments have substantially improved the quality of the paper. D. Paul was supported in part by the NSF grants DMS-1713120, DMS-1811405, DMS-1915894 and CCF-1934568. T. Banerjee and G. Mukherjee were partially supported by NSF DMS-1811866.

Appendix A. Proofs

We first present the detailed proofs of Theorem 1, 3 and Lemma 2. Thereafter, the proofs of all the other results in this paper are provided.

A.1 Preliminary expansions for eigenvector and eigenvalues

In this subsection, we put together the key expansions that are needed to prove the theorems. We first express the j-th sample eigenvector $\hat{\boldsymbol{p}}_j$ as

$$\hat{\boldsymbol{p}}_j = a_j \boldsymbol{P}_K(\boldsymbol{e}_{j,K} + \boldsymbol{\delta}_j) + \sqrt{1-a_j^2}\boldsymbol{P}_{K,\perp}\boldsymbol{u}_{j,n-K}, \tag{17}$$

where $\boldsymbol{P}_K = [\boldsymbol{p}_1 : \cdots : \boldsymbol{p}_K]$, $\boldsymbol{P}_{K,\perp}$ is an $n \times (n-K)$ matrix so that $[\boldsymbol{P}_K : \boldsymbol{P}_{K,\perp}]$ is an orthogonal matrix, $a_j = \|\boldsymbol{P}_K\hat{\boldsymbol{p}}_j\| \in (0,1)$ (without loss of generality, choosing the correct sign), $\boldsymbol{e}_{j,K}$ is the j-th canonical coordinate vector in $\mathbb{R}^K$. Moreover, $\boldsymbol{u}_{j,n-K}$ is uniformly distributed on $\mathbb{S}_{n-K-1}$ (the unit sphere in $\mathbb{R}^{n-K}$), so that $\boldsymbol{u}_{j,n-K} = \boldsymbol{\varepsilon}_j/\sqrt{n-K}$ where $\boldsymbol{\varepsilon}_j \sim N(0, \boldsymbol{I}_{n-K})$. We shall make use of the following asymptotic expansions (Paul, 2007).

$$\|\delta_j\| = O_p(n^{-1/2}) \qquad \text{and} \qquad a_j = \zeta_j + O_p(n^{-1/2}). \tag{18}$$

Now, for $p \le n$, let $\boldsymbol{A}$ be any $p \times n$ matrix such that $\|\boldsymbol{A}\|$ and $\|(\boldsymbol{A}\boldsymbol{A}^T)^{-1}\|$ are bounded even as $p, n \to \infty$. Then, for any $\boldsymbol{b} \in \mathbb{R}^p$ with $\|\boldsymbol{b}\|_2 = 1$, we have the expansion

$$\begin{aligned}
\langle \boldsymbol{b}, \boldsymbol{A}\hat{\boldsymbol{p}}_j\rangle \;=\; & \zeta_j\langle \boldsymbol{b}, \boldsymbol{A}\boldsymbol{p}_j\rangle + \frac{\sqrt{1-\zeta_j^2}}{\sqrt{n-K}}\langle \boldsymbol{b}, \boldsymbol{A}\boldsymbol{P}_{K,\perp}\boldsymbol{\varepsilon}_j\rangle \\
& + (a_j - \zeta_j)\langle \boldsymbol{b}, \boldsymbol{A}\boldsymbol{p}_j\rangle + (\sqrt{1-a_j^2} - \sqrt{1-\zeta_j^2})\frac{1}{\sqrt{n-K}}\langle \boldsymbol{b}, \boldsymbol{A}\boldsymbol{P}_{K,\perp}\boldsymbol{\varepsilon}_j\rangle \\
& + a_j\langle \boldsymbol{b}, \boldsymbol{A}\boldsymbol{P}_K\delta_j\rangle + \sqrt{1-a_j^2}\langle \boldsymbol{b}, \boldsymbol{A}\boldsymbol{P}_{K,\perp}\boldsymbol{\varepsilon}_j\rangle(\|\boldsymbol{\varepsilon}_j\|^{-1} - (n-K)^{-1/2}).
\end{aligned} \tag{19}$$

Suppose that $\mathcal{B}$ be any collection of unit vectors in $\mathbb{R}^p$ of cardinality $O(n^c)$ for some fixed $c \in (0,\infty)$. Then, from equation (19) we conclude that, uniformly over $\boldsymbol{b} \in \mathcal{B}$,

$$\langle \boldsymbol{b}, \boldsymbol{A}\hat{\boldsymbol{p}}_j\rangle - \zeta_j\langle \boldsymbol{b}, \boldsymbol{A}\boldsymbol{p}_j\rangle = O_p(\sqrt{\log n/n}). \tag{20}$$

Here, we used the fact that $\langle \boldsymbol{b}, \boldsymbol{A}\boldsymbol{P}_{K,\perp}\boldsymbol{\varepsilon}_j\rangle \sim N(0, \boldsymbol{b}^T\boldsymbol{A}(\boldsymbol{I}_n - \boldsymbol{P}_K\boldsymbol{P}_K^T)\boldsymbol{A}^T\boldsymbol{b})$, $|\langle \boldsymbol{b}, \boldsymbol{A}\boldsymbol{p}_j\rangle| \le \|\boldsymbol{A}\|$ and, $|\langle \boldsymbol{b}, \boldsymbol{A}\delta_j\rangle| \le \|\boldsymbol{A}\|\|\delta_j\| = O_p(n^{-1/2})$. Moreover, $|a_j - \zeta_j| = O_p(n^{-1/2})$ implies $|\sqrt{1-a_j^2} - \sqrt{1-\zeta_j^2}| = O_p(n^{-1/2})$ and $|\|\boldsymbol{\varepsilon}_j\|^{-1} - (n-K)^{-1/2}| = O_p(n^{-1})$.

A.2 Proof of Theorem 1

First note that for any fixed $(r,\alpha) \in \{-1,0,1\} \times \mathbb{R}$, and any given τ and β, equation (13) gives, for any $\boldsymbol{b} \in \mathcal{B}$ with $\|\boldsymbol{b}\|_2 = 1$

$$\boldsymbol{b}^T\hat{H}_{r,\alpha,\beta}\boldsymbol{b} = \sum_{j=1}^{K}\frac{1}{\hat{\zeta}_j^2}(h_{r,\alpha,\beta}(\hat{\ell}_j^{\mathsf{e}}) - h_{r,\alpha,\beta}(\hat{\ell}_0^{\mathsf{e}}))(\langle \boldsymbol{b}, \hat{\boldsymbol{p}}_j\rangle)^2 + h_{r,\alpha,\beta}(\hat{\ell}_0^{\mathsf{e}})\|\boldsymbol{b}\|^2,$$

and from equations (11), (12), (17) and (18), the above reduces to

$$\begin{aligned} \boldsymbol{b}^T \hat{H}_{r,\alpha,\beta} \boldsymbol{b} &= \sum_{j=1}^{K} (h_{r,\alpha,\beta}(\ell_j) - h_{r,\alpha,\beta}(\ell_0))(\langle \boldsymbol{b}, \boldsymbol{p}_j \rangle)^2 + h_{r,\alpha,\beta}(\ell_0) + O_p(\sqrt{\log n/n}) \\ &= \boldsymbol{b}^T H_{r,\alpha,\beta} \boldsymbol{b} + O_p(\sqrt{\log n/n}), \end{aligned} \tag{21}$$

uniformly over $\boldsymbol{b} \in \mathcal{B}$ consisting of $O(n^c)$ unit vectors, for any fixed $c > 0$. Next, since by assumption **A3**, τ and β belong to compact subsets on which all the quantities in question are smooth functions with uniformly bounded Lipschitz semi-norm with respect to (τ, β), by choosing an appropriate grid of (τ, β) of size $O(n^{c'})$ for some $c' > 0$, we note that the expansion in equation (21) continues to hold uniformly in (τ, β), and hence we have $\sup_{\tau \in \boldsymbol{T}_0, \beta \in \boldsymbol{B}_0, \boldsymbol{b} \in \mathcal{B}} |\boldsymbol{b}^T \hat{H}_{r,\alpha,\beta} \boldsymbol{b} - \boldsymbol{b}^T H_{r,\alpha,\beta} \boldsymbol{b}| = O_p(\sqrt{\log n/n})$, thus proving the theorem.

A.3 Proof of Lemma 2

First note that,

$$\hat{H}^{\mathsf{un}}_{r,\alpha,\beta} = \sum_{j=1}^{K} d_{r,\alpha,\beta}(\hat{\ell}_j) \hat{\boldsymbol{p}}_j \hat{\boldsymbol{p}}_j^T + d_{r,\alpha,\beta}(\hat{\ell}_0)(\boldsymbol{I}_n - \sum_{j=1}^{K} \hat{\boldsymbol{p}}_j \hat{\boldsymbol{p}}_j^T),$$

where $d_{r,\alpha,\beta}(x) = (x^{-1} + \tau^{-1} x^{-\beta})^{-r} x^{\alpha}$. Thus, $\boldsymbol{b}^T \hat{H}^{\mathsf{un}}_{r,\alpha,\beta} \boldsymbol{b}$ equals

$$\sum_{j=1}^{K} d_{r,\alpha,\beta}(\hat{\ell}_j)(\boldsymbol{b}^T \hat{\boldsymbol{p}}_j)^2 + d_{r,\alpha,\beta}(\hat{\ell}_0) \left[\boldsymbol{b}^T \boldsymbol{b} - \sum_{j=1}^{K} (\boldsymbol{b}^T \hat{\boldsymbol{p}}_j)^2 \right] . \tag{22}$$

From equation (9) it follows that for all $j = 1, \ldots, K$,

$$d_{r,\alpha,\beta}(\hat{\ell}_j) = d_{r,\alpha,\beta}\big(\ell_0 \psi(\ell_j/\ell_0, \rho)\big) + O_p(n^{-1/2}) \text{ and,} \tag{23}$$

$$d_{r,\alpha,\beta}(\hat{\ell}_0) = d_{r,\alpha,\beta}(\ell_0) + O_p(n^{-1/2}). \tag{24}$$

On the other hand, using (10) it follows for $j = 1, \ldots, K$ that

$$\boldsymbol{b}^T \hat{\boldsymbol{p}}_j = \zeta_j \boldsymbol{b}^T \boldsymbol{p}_j + \sqrt{1 - \zeta_j^2} \boldsymbol{b}^T (\boldsymbol{I}_n - \boldsymbol{P}_K \boldsymbol{P}_K^T) \frac{\boldsymbol{\varepsilon}_j}{\sqrt{n - K}} + O_p(n^{-1/2}) \ ,$$

which implies that

$$(\boldsymbol{b}^T \hat{\boldsymbol{p}}_j)^2 = \zeta_j^2 (\boldsymbol{b}^T \hat{\boldsymbol{p}}_j)^2 + O_p(n^{-1/2}). \tag{25}$$

Now, substituting equations (23) to (25) in equation (22) we have

$$\boldsymbol{b}^T \hat{H}^{\mathsf{un}}_{r,\alpha,\beta} \boldsymbol{b} = \sum_{j=1}^{K} \big\{ d_{r,\alpha,\beta}\big(\ell_0 \psi(\ell_j/\ell_0, \rho)\big) - d_{r,\alpha,\beta}(\ell_0) \big\} \zeta_j^2 (\boldsymbol{b}^T \hat{\boldsymbol{p}}_j)^2 + d_{r,\alpha,\beta}(\ell_0) \boldsymbol{b}^T \boldsymbol{b} \ .$$

Also, we know that

$$\boldsymbol{b}^T H_{r,\alpha,\beta} \boldsymbol{b} = \sum_{j=1}^{K} \big\{ d_{r,\alpha,\beta}\big(\ell_j\big) - d_{r,\alpha,\beta}(\ell_0) \big\} (\boldsymbol{b}^T \hat{\boldsymbol{p}}_j)^2 + d_{r,\alpha,\beta}(\ell_0) \boldsymbol{b}^T \boldsymbol{b} \ .$$

Comparing the above two displays, the result follows.

A.4 Proof of Theorem 3

We only prove the result for fixed (τ, β) since the argument can be extended to compact subsets of (τ, β), under assumption **A3**, using an argument similar to that used in the proof of Theorem 1.

Since the aggregated Bayes predictive rules involve quadratic forms of the form $\boldsymbol{b}^T G_{r,\alpha,\beta}\boldsymbol{b}$, we have the following cases of interest: $G_{0,1,0}, G_{1,0,\beta}$ and $G_{1,-1,\beta}$. In order to analyze the corresponding estimators of these quantities of interest, we introduce some notations. Let $\boldsymbol{C} = \boldsymbol{A}\boldsymbol{A}^T$, $\boldsymbol{Q} = [\boldsymbol{q}_1 : \cdots : \boldsymbol{q}_K]$, where $\boldsymbol{q}_j = \boldsymbol{A}\boldsymbol{p}_j$, and $\widetilde{\boldsymbol{Q}} = [\widetilde{\boldsymbol{q}}_1 : \cdots : \widetilde{\boldsymbol{q}}_K]$ where $\widetilde{\boldsymbol{q}}_j = \hat{\zeta}_j^{-1}\boldsymbol{A}\hat{\boldsymbol{p}}_j$. Then, for any $\beta \in \mathbb{R}^+$,

$$\begin{aligned}
\boldsymbol{A}\hat{H}_{0,\beta,0}\boldsymbol{A}^T &= (\hat{\ell}_0^{\mathsf{e}})^\beta \boldsymbol{A}\boldsymbol{A}^T + \sum_{j=1}^K \frac{(\hat{\ell}_j^{\mathsf{e}})^\beta - (\hat{\ell}_0^{\mathsf{e}})^\beta}{\hat{\zeta}_j^2}\boldsymbol{A}\hat{\boldsymbol{p}}_j\hat{\boldsymbol{p}}_j^T\boldsymbol{A}^T \\
&= (\hat{\ell}_0^{\mathsf{e}})^\beta \boldsymbol{C} + \sum_{j=1}^K \left((\hat{\ell}_j^{\mathsf{e}})^\beta - (\hat{\ell}_0^{\mathsf{e}})^\beta\right)\widetilde{\boldsymbol{q}}_j\widetilde{\boldsymbol{q}}_j^T \\
&= (\hat{\ell}_0^{\mathsf{e}})^\beta \left[\boldsymbol{C} + \widetilde{\boldsymbol{Q}}\left((\hat{\ell}_0^{\mathsf{e}})^{-\beta}\hat{\Lambda}^\beta - \boldsymbol{I}_K\right)\widetilde{\boldsymbol{Q}}^T\right], \qquad (26)
\end{aligned}$$

where, $\hat{\Lambda} = \mathsf{diag}(\hat{\ell}_1^{\mathsf{e}}, \ldots, \hat{\ell}_K^{\mathsf{e}})$. Setting $\beta = 1$, we observe that $\boldsymbol{b}^T\hat{G}_{0,1,0}\boldsymbol{b} = \hat{\ell}_0^{\mathsf{e}}\boldsymbol{b}^T\boldsymbol{A}\boldsymbol{A}^T\boldsymbol{b} + \sum_{j=1}^K(\hat{\ell}_j^{\mathsf{e}} - \hat{\ell}_0^{\mathsf{e}})\frac{1}{\hat{\zeta}_j^2}\left(\langle\boldsymbol{b}, \boldsymbol{A}\hat{\boldsymbol{p}}_j\rangle\right)^2$ which is $\boldsymbol{b}^T G_{0,1,0}\boldsymbol{b} + O_p(\sqrt{\log n/n})$ from equations (20) and (18). This proves the theorem when $\beta = 1$.

To prove the theorem for any $\beta \neq 1$, we make repeated use of the following basic formula for matrix inversion. Given a symmetric nonsingular $p \times p$ matrix $\boldsymbol{B}$, and a $p \times q$ matrix $\boldsymbol{D}$,

$$\left(\boldsymbol{B} + \boldsymbol{D}\boldsymbol{D}^T\right)^{-1} = \boldsymbol{B}^{-1} - \boldsymbol{B}^{-1}\boldsymbol{D}\left(\boldsymbol{I}_q + \boldsymbol{D}^T\boldsymbol{B}^{-1}\boldsymbol{D}\right)^{-1}\boldsymbol{D}^T\boldsymbol{B}^{-1}. \qquad (27)$$

Using equations (27) and (26), we have, with $\hat{\Delta}_\beta = (\hat{\ell}_0^{\mathsf{e}})^{-\beta}\hat{\Lambda}^\beta - \boldsymbol{I}_K$,

$$(\boldsymbol{A}\hat{H}_{0,\beta,0}\boldsymbol{A}^T)^{-1} = (\hat{\ell}_0^{\mathsf{e}})^{-\beta}\boldsymbol{C}^{-1} - (\hat{\ell}_0^{\mathsf{e}})^{-\beta}\boldsymbol{C}^{-1}\widetilde{\boldsymbol{Q}}\hat{\Delta}_\beta^{1/2}\left[\boldsymbol{I}_K + \hat{\Delta}_\beta^{1/2}\widetilde{\boldsymbol{Q}}^T\boldsymbol{C}^{-1}\widetilde{\boldsymbol{Q}}\hat{\Delta}_\beta^{1/2}\right]^{-1}\hat{\Delta}_\beta^{1/2}\widetilde{\boldsymbol{Q}}^T\boldsymbol{C}^{-1}.$$

We must, therefore, analyze the behavior of $\widetilde{\boldsymbol{Q}}^T\boldsymbol{C}^{-1}\widetilde{\boldsymbol{Q}}$. As a preliminary step, we observe that since $\boldsymbol{A}\boldsymbol{P}_{K,\perp}\boldsymbol{\varepsilon}_j \sim N(0, \boldsymbol{A}(\boldsymbol{I}_n - \boldsymbol{P}_K\boldsymbol{P}_K^T)\boldsymbol{A}^T)$, it follows that

$$\frac{1}{p}\|\boldsymbol{C}^{-1/2}\boldsymbol{A}\boldsymbol{P}_{K,\perp}\boldsymbol{\varepsilon}_j\|^2 = \frac{1}{p}\mathrm{trace}\left(\boldsymbol{C}^{-1}\boldsymbol{A}(\boldsymbol{I}_n - \boldsymbol{P}_K\boldsymbol{P}_K^T)\boldsymbol{A}^T\right) + O_p(p^{-1/2}), \qquad (28)$$

which reduces to $1 - r_A/p + O_p(p^{-1/2})$ where $r_A = p^{-1}\mathrm{trace}\left(\boldsymbol{A}^T\boldsymbol{C}^{-1}\boldsymbol{A}\boldsymbol{P}_K\boldsymbol{P}_K^T\right)$ and

$$|r_A| \leq \|\boldsymbol{A}^T\boldsymbol{C}^{-1}\boldsymbol{A}\| \cdot \mathrm{rank}(\boldsymbol{P}_K\boldsymbol{P}_K^T) \leq K.$$

We will use equation (28) and the bound on $|r_A|$ to control $\widetilde{\boldsymbol{Q}}^T\boldsymbol{C}^{-1}\widetilde{\boldsymbol{Q}}$. First note that using equations (17), (18) and, for any $1 \leq j, k \leq K$,

$$\begin{aligned}
\widetilde{\boldsymbol{q}}_j^T\boldsymbol{C}^{-1}\widetilde{\boldsymbol{q}}_k &= \boldsymbol{p}_j^T\boldsymbol{A}^T\boldsymbol{C}^{-1}\boldsymbol{A}\boldsymbol{p}_k(1 + O_p(n^{-1/2})) \\
&\quad + \frac{\sqrt{1-\zeta_j^2}\sqrt{1-\zeta_k^2}}{\zeta_j\zeta_k}\frac{\langle\boldsymbol{A}\boldsymbol{P}_{K,\perp}\boldsymbol{\varepsilon}_k, \boldsymbol{C}^{-1}\boldsymbol{A}\boldsymbol{P}_{K,\perp}\boldsymbol{\varepsilon}_j\rangle}{n-K}(1 + O_p(n^{-1/2})),
\end{aligned}$$

which, using equation (28), reduces to

$$\boldsymbol{p}_j^T \boldsymbol{A}^T \boldsymbol{C}^{-1} \boldsymbol{A} \boldsymbol{p}_k (1 + O_p(n^{-1/2})) + \frac{\sqrt{1-\zeta_j^2}\sqrt{1-\zeta_k^2}}{\zeta_j \zeta_k} \frac{p - r_A + O_p(p^{1/2})}{n-K} (1 + O_p(n^{-1/2})),$$

and finally to $\boldsymbol{q}_j^T \boldsymbol{C}^{-1} \boldsymbol{q}_k (1 + O_p(n^{-1/2})) + O_p(p/n)$ using the bound on $|r_A|$. Consequently, we have

$$\widetilde{\boldsymbol{Q}}^T \boldsymbol{C}^{-1} \widetilde{\boldsymbol{Q}} = \boldsymbol{Q}^T \boldsymbol{C}^{-1} \boldsymbol{Q} + O_p(n^{-1/2}) + O_p(p/n). \tag{29}$$

Now let $\Delta_\beta = \ell_0^{-\beta} \Lambda - \boldsymbol{I}_K$. Then $\|\hat{\Delta}_\beta - \Delta_\beta\| = O_p(n^{-1/2})$, and hence, by equation (29),

$$\hat{\boldsymbol{U}}_\beta^{-1} := \left[\boldsymbol{I}_K + \hat{\Delta}_\beta^{1/2} \widetilde{\boldsymbol{Q}}^T \boldsymbol{C}^{-1} \widetilde{\boldsymbol{Q}} \hat{\Delta}_\beta^{1/2}\right]^{-1} = \left[\boldsymbol{I}_K + \Delta_\beta^{1/2} \boldsymbol{Q}^T \boldsymbol{C}^{-1} \boldsymbol{Q} \Delta_\beta^{1/2}\right]^{-1} + \boldsymbol{R}_{\beta,n} = \boldsymbol{U}_\beta^{-1} + \boldsymbol{R}_{\beta,n}, \tag{30}$$

where $\|\boldsymbol{R}_{\beta,n}\| = O_p(n^{-1/2}) + O_p(p/n)$. Furthermore, $(\boldsymbol{A}\hat{H}_{0,1,0}\boldsymbol{A}^T)^{-1} + \tau^{-1}(\boldsymbol{A}\hat{H}_{0,\beta,0}\boldsymbol{A}^T)^{-1}$ can be written as

$$\left((\hat{\ell}_0^{\mathsf{e}})^{-1} + \frac{1}{\tau}(\hat{\ell}_0^{\mathsf{e}})^{-\beta}\right) \boldsymbol{C}^{-1} - \boldsymbol{C}^{-1} \widetilde{\boldsymbol{Q}} \left[(\hat{\ell}_0^{\mathsf{e}})^{-1} \hat{\Delta}_1^{1/2} \hat{\boldsymbol{U}}_1 \hat{\Delta}_1^{1/2} + \frac{1}{\tau}(\hat{\ell}_0^{\mathsf{e}})^{-\beta} \hat{\Delta}_\beta^{1/2} \hat{\boldsymbol{U}}_\beta \hat{\Delta}_\beta^{1/2}\right] \widetilde{\boldsymbol{Q}}^T \boldsymbol{C}^{-1},$$

which by equation (30) is

$$\left((\hat{\ell}_0^{\mathsf{e}})^{-1} + \frac{1}{\tau}(\hat{\ell}_0^{\mathsf{e}})^{-\beta}\right) \boldsymbol{C}^{-1} - \boldsymbol{C}^{-1} \widetilde{\boldsymbol{Q}} \hat{\boldsymbol{V}}_\beta \widetilde{\boldsymbol{Q}}^T \boldsymbol{C}^{-1}, \tag{31}$$

where $\hat{\boldsymbol{V}}_\beta = \boldsymbol{V}_\beta + \check{\boldsymbol{R}}_{1,n} + \frac{1}{\tau}\check{\boldsymbol{R}}_{\beta,n}$ with $\boldsymbol{V}_\beta = \ell_0^{-1}\Delta_1^{1/2}\boldsymbol{U}_1\Delta_1^{1/2} + \frac{1}{\tau}\ell_0^{-\beta}\Delta_\beta^{1/2}\boldsymbol{U}_\beta\Delta_\beta^{1/2}$ and $\check{\boldsymbol{R}}_{\beta,n} = \hat{\ell}_0^{-\beta}\hat{\Delta}_\beta^{1/2}\boldsymbol{R}_{\beta,n}\hat{\Delta}_\beta^{1/2}$, so that $\|\check{\boldsymbol{R}}_{\beta,n}\| = O_p(n^{-1/2}) + O_p(p/n)$ for all β. Notice that $\boldsymbol{V}_\beta$ is positive definite, and hence $\hat{\boldsymbol{V}}_\beta$ is positive definite with probability tending to 1.

Define, for $x > 0$, $a_{\beta,\tau}(x) = x^{-1} + \tau^{-1}x^{-\beta}$. By equation (31), we can write

$$\left[(\boldsymbol{A}\hat{H}_{0,1,0}\boldsymbol{A}^T)^{-1} + \tau^{-1}(\boldsymbol{A}\hat{H}_{0,\beta,0}\boldsymbol{A}^T)^{-1}\right]^{-1} = \frac{1}{a_{\beta,\tau}(\hat{\ell}_0^{\mathsf{e}})} \boldsymbol{C} \left[\boldsymbol{C} - \frac{1}{a_{\beta,\tau}(\hat{\ell}_0^{\mathsf{e}})} \widetilde{\boldsymbol{Q}} \hat{\boldsymbol{V}}_\beta \widetilde{\boldsymbol{Q}}^T\right]^{-1} \boldsymbol{C}$$

$$= \frac{1}{a_{\beta,\tau}(\hat{\ell}_0^{\mathsf{e}})} \boldsymbol{C} + \frac{1}{(a_{\beta,\tau}(\hat{\ell}_0^{\mathsf{e}}))^2} \widetilde{\boldsymbol{Q}} \left[\hat{\boldsymbol{V}}_\beta^{-1} - \frac{1}{a_{\beta,\tau}(\hat{\ell}_0^{\mathsf{e}})} \widetilde{\boldsymbol{Q}}^T \boldsymbol{C}^{-1} \widetilde{\boldsymbol{Q}}\right]^{-1} \widetilde{\boldsymbol{Q}}^T,$$

and using $\hat{\boldsymbol{V}}_\beta = \boldsymbol{V}_\beta + \check{\boldsymbol{R}}_{1,n} + \frac{1}{\tau}\check{\boldsymbol{R}}_{\beta,n}$, we can re-write it as

$$\frac{1}{a_{\beta,\tau}(\hat{\ell}_0^{\mathsf{e}})} \boldsymbol{C} + \frac{1}{(a_{\beta,\tau}(\hat{\ell}_0^{\mathsf{e}}))^2} \widetilde{\boldsymbol{Q}} \left[\boldsymbol{V}_\beta^{-1} - \frac{1}{a_{\beta,\tau}(\ell_0)} \boldsymbol{Q}^T \boldsymbol{C}^{-1} \boldsymbol{Q} + \boldsymbol{R}_{*,n}\right]^{-1} \widetilde{\boldsymbol{Q}}^T,$$

where $\|\boldsymbol{R}_{*,n}\| = O_p(n^{-1/2}) + O_p(p/n)$. As a consequence, we have

$$\boldsymbol{b}^T \hat{G}_{1,0,\beta} \boldsymbol{b} = \boldsymbol{b}^T G_{1,0,\beta} \boldsymbol{b} + O_p(\sqrt{\log n/n}) + O_p(p/n),$$

uniformly over $\boldsymbol{b} \in \mathcal{B}$. An analogous calculation yields

$$\boldsymbol{b}^T \hat{G}_{1,-1,\beta} \boldsymbol{c} = \boldsymbol{b}^T G_{1,-1,\beta} \boldsymbol{c} + O_p(\sqrt{\log n/n}) + O_p(p/n),$$

uniformly over $\boldsymbol{b}, \boldsymbol{c} \in \mathcal{B}$.

A.5 Proof of Lemma 1

First note that under the hierarchical model of equations (1) and (2), the posterior distribution of $\boldsymbol{\psi}$ given $\boldsymbol{AX}$ is $N(\boldsymbol{A\eta}_0 + G_{1,-1,\beta}\boldsymbol{A}(\boldsymbol{X}-\boldsymbol{\eta}_0), G_{1,0,\beta})$. To prove this Lemma, we first fix a few notations. For coordinate i, let $v_i, v_{\psi i}$ and v_{fi} denote the i^{th} diagonal element of $\check{\boldsymbol{\Sigma}}_1$, $\check{\boldsymbol{\Sigma}}_\beta$ and $m_0^{-1}\check{\boldsymbol{\Sigma}}_1$, respectively. Suppose $(\boldsymbol{AX})_i$ and $(\boldsymbol{A\eta}_0)_i$ be the ith coordinates of $\boldsymbol{AX}$ and $\boldsymbol{A\eta}_0$, respectively. The minimizer of the univariate Bayes risk $B_i(\tau,\beta)$ is given by

$$\hat{q}_i = \arg\min_q \int L_i(\psi_i, q_i)\pi(\psi_i|(\boldsymbol{AX})_i),$$

where the posterior distribution $\pi(\psi_i|(\boldsymbol{AX})_i) \sim N(\gamma_i, \omega_i)$ with $\gamma_i = \delta_i(\boldsymbol{AX})_i + (1-\delta_i)(\boldsymbol{A\eta}_0)_i$, $\delta_i = v_{\psi i}/(v_{\psi i}+v_i)$ and $\omega_i = (v_{\psi i}^{-1}+v_i^{-1})^{-1}$. We prove the Lemma for the generalized absolute loss and the Linex loss functions. The univariate Bayes predictive rules for the other losses considered in this paper will follow from similar arguments.

For the Linex loss function, note that $L_i(\psi_i, q_i) = \mathbb{E}_{V_i}\mathcal{L}_i(V_i, q_i)$ where $\mathcal{L}_i(V_i, q_i)$ is the Linex loss for coordinate i from equation (4). Since $V_i \sim N(\psi_i, v_{fi})$,

$$\mathbb{E}_{V_i}\mathcal{L}_i(V_i, q_i) = b_i\Big[\exp\{a_i(q_i-\psi_i) + (a_i^2/2)v_{fi}\} - a_i(q_i-\psi_i) - 1\Big].$$

Furthermore, $\mathbb{E}_{\psi_i|(\boldsymbol{AX})_i}L_i(\psi_i, q_i) = b_i\Big[\exp\{a_i(q_i-\gamma_i) + (a_i^2/2)(v_{fi}+\omega_i)\} - a_i(q_i-\gamma_i) - 1\Big]$ is convex in q_i. Differentiating the above posterior expectation with respect to q_i, we get

$$\hat{q}_i = \delta_i(\boldsymbol{AX})_i + (1-\delta_i)(\boldsymbol{A\eta}_0)_i - \frac{a_i}{2}(v_{fi}+\omega_i),$$

which completes the proof.

For the generalized absolute loss function in equation (3), note that

$$\mathbb{E}_{V_i}\mathcal{L}_i(V_i, q_i) = b_i(\psi_i - q_i) + (b_i+h_i)\mathbb{E}(q_i-\psi_i-Z)^+,$$

where Z is a standard normal random variable. Furthermore, direct calculation yields $\mathbb{E}(q_i-\psi_i-Z)^+ = (q_i-\psi_i)\Phi(q_i-\psi_i) + \phi(q_i-\psi_i)$. The Bayes predictive rule then follows from Proposition 2.1 and 2.2 of Mukherjee et al. (2015).

A.6 Proof of Lemma 3

We prove this lemma for the generalized absolute loss function in equation (3). For any i and fixed (τ,β), it follows from Theorem 1,

$$\Big|\hat{q}_i^{\mathsf{S}}(\boldsymbol{X}|\boldsymbol{S},\tau,\beta) - q_i^{\mathsf{Bayes}}(\boldsymbol{X}|\boldsymbol{\Sigma},\tau,\beta)\Big| \le \Big|\boldsymbol{e}_i^T(\hat{H}_{1,-1,\beta} - H_{1,-1,\beta})(\boldsymbol{X}-\boldsymbol{\eta}_0)\Big| + O_p\Big(\sqrt{\frac{\log n}{n}}\Big).$$

The first term on the right of the inequality above can be written as a difference of two symmetric quadratic forms as follows

$$\frac{\|\boldsymbol{X}-\boldsymbol{\eta}_0\|_2}{4}\Big[\Big|(\boldsymbol{a}+\boldsymbol{e}_i)^T(\hat{H}_{1,-1,\beta} - H_{1,-1,\beta})(\boldsymbol{a}+\boldsymbol{e}_i)\Big| - \Big|(\boldsymbol{a}-\boldsymbol{e}_i)^T(\hat{H}_{1,-1,\beta} - H_{1,-1,\beta})(\boldsymbol{a}-\boldsymbol{e}_i)\Big|\Big],$$

where $\boldsymbol{a} = (\boldsymbol{X}-\boldsymbol{\eta}_0)/\|\boldsymbol{X}-\boldsymbol{\eta}_0\|_2$.

Now, re-apply Theorem 1 separately to these two quadratic forms and note that the above is bounded by $O_p(\sqrt{\log n/n})(\|\boldsymbol{X}-\boldsymbol{\eta}_0\|_2)(\|\boldsymbol{a}+\boldsymbol{e}_i\|_2^2 + \|\boldsymbol{a}-\boldsymbol{e}_i\|_2^2)/4$, from which the result follows.

A.7 Proofs of Lemmata 4 and 5

To prove these lemmata, we use the following result.

Lemma 6. *Under assumptions **A1** and **A2**, uniformly in $\boldsymbol{b} \in \mathcal{B}$ such that $\mathcal{B} = O(n^c)$ for any fixed $c > 0$, with $\|\boldsymbol{b}\|_2 = 1$, and for all $(r, \alpha) \in \{-1, 0, 1\} \times \mathbb{R}$, we have as $n \to \infty$,*

$$\sup_{\boldsymbol{b}\in\mathcal{B}} \Big| \boldsymbol{b}^T \hat{H}_{1,-1,\beta} \mathcal{J}(\boldsymbol{\Sigma}) \hat{H}_{1,-1,\beta} \boldsymbol{b} - \boldsymbol{b}^T H_{1,-1,\beta} \mathcal{J}(\boldsymbol{\Sigma}) H_{1,-1,\beta} \boldsymbol{b} - j(\ell_0) \sum_{j=1}^{K} (h_{1,-1,\beta}(\ell_j) - h_{1,-1,\beta}(\ell_0))^2 (\langle \boldsymbol{b}, \boldsymbol{p}_j \rangle)^2 \Big| = O_p(\sqrt{\log n / n}) \ .$$

Proof of Lemma 6. Let us first define the following quantities: $\Delta_j(h_1) = h_{1,-1,\beta}(\ell_j) - h_{1,-1,\beta}(\ell_0)$, $\Delta_j = \mathcal{J}(\ell_j) - \mathcal{J}(\ell_0)$ and $\hat{\Delta}_j(h_1) = h_{1,-1,\beta}(\hat{\ell}_j^{\mathsf{e}}) - h_{1,-1,\beta}(\hat{\ell}_0^{\mathsf{e}})$ where $h_{r,\alpha,\beta}$ is the scalar version of $H_{r,\alpha,\beta}$ and $\mathcal{J}(x) = x + \tau x^\beta$ being the scalar version of $\mathcal{J}(\boldsymbol{\Sigma})$. For any $\boldsymbol{b} \in \mathcal{B}$ with $\|\boldsymbol{b}\|_2 = 1$, expand $\boldsymbol{b}^T \hat{H}_{1,-1,\beta} \mathcal{J}(\boldsymbol{\Sigma}) \hat{H}_{1,-1,\beta}^T \boldsymbol{b}$ as

$$\begin{aligned}
&\sum_{j=1}^{K}\sum_{j'=1}^{K}\sum_{k=1}^{K} \frac{\hat{\Delta}_j(h)\hat{\Delta}_{j'}(h)}{\hat{\zeta}_j^2 \hat{\zeta}_{j'}^2} \Delta_k \langle \boldsymbol{b}, \hat{\boldsymbol{p}}_j \rangle \langle \boldsymbol{b}, \hat{\boldsymbol{p}}_{j'} \rangle \langle \boldsymbol{p}_k, \hat{\boldsymbol{p}}_j \rangle \langle \boldsymbol{p}_k, \hat{\boldsymbol{p}}_{j'} \rangle \\
&+ \ \mathcal{J}(\ell_0) \sum_{j=1}^{K}\sum_{j'=1}^{K} \frac{\hat{\Delta}_j(h)\hat{\Delta}_{j'}(h)}{\hat{\zeta}_j^2 \hat{\zeta}_{j'}^2} \langle \boldsymbol{b}, \hat{\boldsymbol{p}}_j \rangle \langle \boldsymbol{b}, \hat{\boldsymbol{p}}_{j'} \rangle \langle \hat{\boldsymbol{p}}_{j'}, \hat{\boldsymbol{p}}_j \rangle \\
&+ \ 2h(\hat{\ell}_0^{\mathsf{e}}) \sum_{j=1}^{K}\sum_{k=1}^{K} \frac{\hat{\Delta}_j(h)}{\hat{\zeta}_j^2} \Delta_k \langle \boldsymbol{b}, \hat{\boldsymbol{p}}_j \rangle \langle \boldsymbol{b}, \boldsymbol{p}_k \rangle \langle \boldsymbol{p}_k, \hat{\boldsymbol{p}}_j \rangle + 2h(\hat{\ell}_0^{\mathsf{e}}) \mathcal{J}(\ell_0) \sum_{j=1}^{K} \frac{\hat{\Delta}_j(h)}{\hat{\zeta}_j^2} (\langle \boldsymbol{b}, \hat{\boldsymbol{p}}_j \rangle)^2 \\
&+ \ (h(\hat{\ell}_0^{\mathsf{e}}))^2 \sum_{k=1}^{K} \Delta_k (\langle \boldsymbol{b}, \boldsymbol{p}_k \rangle)^2 + (h(\hat{\ell}_0^{\mathsf{e}}))^2 \mathcal{J}(\ell_0) \|\boldsymbol{b}\|^2.
\end{aligned}$$

Then, using equation (9), it can be verified that above asymptotically equals

$$\begin{aligned}
&\boldsymbol{b}^T \Big(\sum_{j=1}^{K} \Delta_j(h_1) \boldsymbol{p}_j \boldsymbol{p}_j^T + h_{1,-1,\beta}(\ell_0) \boldsymbol{I}_n \Big) \Big(\sum_{j=1}^{K} \Delta_j \boldsymbol{p}_j \boldsymbol{p}_j^T + h_{1,-1,\beta}(\ell_0) \boldsymbol{I}_n \Big) \\
&\Big(\sum_{j=1}^{K} \Delta_j(h_1) \boldsymbol{p}_j \boldsymbol{p}_j^T + h_{1,-1,\beta}(\ell_0) \boldsymbol{I}_n \Big) \boldsymbol{b} + \ \mathcal{J}(\ell_0) \sum_{j=1}^{K} \Big(\Delta_j(h_1) \Big)^2 (\langle \boldsymbol{b}, \boldsymbol{p}_j \rangle)^2 + O_p(\sqrt{\log n / n}),
\end{aligned}$$

where the O_p term is uniform in $\boldsymbol{b} \in \mathcal{B}$ consisting of $O(n^c)$ unit vectors. Finally, using the definitions of $\Delta_j(h_1)$, Δ_j and arguments similar that used in proving Theorem 1, the result follows.

Next, we prove the three statements of Lemma 4.

Proof of Lemma 4, statement (a) - first note that $\mathbb{E}\Big\{ \Big(q_i^{\mathsf{cs}}(\boldsymbol{X}|\boldsymbol{S}, f_i, \tau, \beta) - q_i^{\mathsf{Bayes}}(\boldsymbol{X}|\boldsymbol{\Sigma}, \tau, \beta) \Big)^2 \Big\}$ can be decomposed as

$$\mathbb{E}^2 \Big\{ \Big(q_i^{\mathsf{cs}}(\boldsymbol{X}|\boldsymbol{S}, f_i) - q_i^{\mathsf{Bayes}}(\boldsymbol{X}) \Big) \Big\} + \mathrm{Var} \Big\{ \Big(q_i^{\mathsf{cs}}(\boldsymbol{X}|\boldsymbol{S}, f_i) - q_i^{\mathsf{Bayes}}(\boldsymbol{X}) \Big) \Big\},$$

where the first term represents bias squared and the second term is the variance. Now consider, for example, the generalized absolute loss function of equation (3). Under this loss, the bias with respect to the marginal distribution of $\boldsymbol{X}$ is

$$\Phi^{-1}(\tilde{b}_i)\Big[\Big(\boldsymbol{e}_i^T\hat{H}_{1,0,\beta}\boldsymbol{e}_i + m_0^{-1}\boldsymbol{e}_i^T\hat{H}_{0,1,0}\boldsymbol{e}_i\Big)^{1/2} - \Big(\boldsymbol{e}_i^T H_{1,0,\beta}\boldsymbol{e}_i + m_0^{-1}\boldsymbol{e}_i^T H_{0,1,0}\boldsymbol{e}_i\Big)^{1/2}\Big],$$

which, by Theorem 1 is $O_p(\sqrt{\log n/n})$. Now the variance term is equal to

$$f_i^2\boldsymbol{e}_i^T\hat{H}_{1,-1,\beta}\mathcal{J}(\boldsymbol{\Sigma})\hat{H}_{1,-1,\beta}^T\boldsymbol{e}_i - 2f_i\boldsymbol{e}_i^T\hat{H}_{1,-1,\beta}\mathcal{J}(\boldsymbol{\Sigma})H_{1,-1,\beta}^T\boldsymbol{e}_i + \boldsymbol{e}_i^T H_{1,-1,\beta}\mathcal{J}(\boldsymbol{\Sigma})H_{1,-1,\beta}^T\boldsymbol{e}_i\ ,$$

which is a quadratic with respect to f_i and is minimized at

$$f_i^{\mathsf{OR}} = \frac{\boldsymbol{e}_i^T\hat{H}_{1,-1,\beta}\mathcal{J}(\boldsymbol{\Sigma})H_{1,-1,\beta}^T\boldsymbol{e}_i}{\boldsymbol{e}_i^T\hat{H}_{1,-1,\beta}\mathcal{J}(\boldsymbol{\Sigma})\hat{H}_{1,-1,\beta}^T\boldsymbol{e}_i}\ .$$

The numerator in the above expression is an asymmetric quadratic form in $\hat{H}_{r,\alpha,\beta}$ and by Theorem 1 it equals $\boldsymbol{e}_i^T H_{1,-1,\beta}\mathcal{J}(\boldsymbol{\Sigma})H_{1,-1,\beta}^T\boldsymbol{e}_i + (\boldsymbol{e}_i^T H_{1,-1,\beta}\mathcal{J}^2(\boldsymbol{\Sigma})H_{1,-1,\beta}^T\boldsymbol{e}_i)^{1/2}O_p(\sqrt{\log n/n})$. By Lemma 6, the denominator is

$$\boldsymbol{e}_i^T H_{1,-1,\beta}\mathcal{J}(\boldsymbol{\Sigma})H_{1,-1,\beta}\boldsymbol{e}_i + j(\ell_0)\sum_{j=1}^{K}(h_{1,-1,\beta}(\ell_j) - h_{1,-1,\beta}(\ell_0))^2(\langle\boldsymbol{e}_i,\boldsymbol{p}_j\rangle)^2 + O_p(\sqrt{\log n/n}),$$

which, for fixed $\tau > 0, \beta \geq 0$, is non-trivial since $\ell_j > \ell_0 > 0$ for all $j = 1,\ldots,K$. Thus, the ratio asymptotically equals

$$\frac{\boldsymbol{e}_i^T H_{1,-1,\beta}\mathcal{J}(\boldsymbol{\Sigma})H_{1,-1,\beta}^T\boldsymbol{e}_i}{\boldsymbol{e}_i^T H_{1,-1,\beta}\mathcal{J}(\boldsymbol{\Sigma})H_{1,-1,\beta}^T\boldsymbol{e}_i + j(\ell_0)\sum_{j=1}^{K}(h_{1,-1,\beta}(\ell_j) - h_{1,-1,\beta}(\ell_0))^2(\langle\boldsymbol{e}_i,\boldsymbol{p}_j\rangle)^2} + O_p\Big(\sqrt{\frac{\log n}{n}}\Big).$$

The second term in the denominator is at least as big as

$$j(\ell_0)(h_{1,-1,\beta}(\ell_K) - h_{1,-1,\beta}(\ell_0))^2\|\boldsymbol{P}_K\boldsymbol{e}_i\|^2.$$

Finally, note that $U(\boldsymbol{\Sigma}) = H_{1,-1,\beta}\mathcal{J}(\boldsymbol{\Sigma})H_{1,-1,\beta}$, from which the result follows.

Proof of Lemma 4, statement (b) - from Lemma 6, $\boldsymbol{b}^T\hat{H}_{1,-1,\beta}\mathcal{J}(\boldsymbol{\Sigma})\hat{H}_{1,-1,\beta}^T\boldsymbol{b}$ asymptotically equals

$$\boldsymbol{b}^T H_{1,-1,\beta}\mathcal{J}(\boldsymbol{\Sigma})H_{1,-1,\beta}\boldsymbol{b} + j(\ell_0)\sum_{j=1}^{K}(h_{1,-1,\beta}(\ell_j) - h_{1,-1,\beta}(\ell_0))^2(\langle\boldsymbol{b},\boldsymbol{p}_j\rangle)^2 + O_p(\sqrt{\log n/n}),$$

which is strictly bigger than $\boldsymbol{b}^T H_{1,-1,\beta}\mathcal{J}(\boldsymbol{\Sigma})H_{1,-1,\beta}\boldsymbol{b} + O_p(\sqrt{\log n/n})$ for any fixed $\tau > 0, \beta > 0$ and from this the proof immediately follows.

Proof of Lemma 4, statement (c) - this follows directly from statements (a) and (b). For any coordinate i, by definition of f_i^{OR} in statement (a),

$$\mathbb{E}\Big[\Big(q_i^{\mathsf{S}}(\boldsymbol{X}|\boldsymbol{S}) - q_i^{\mathsf{Bayes}}(\boldsymbol{X})\Big)^2\Big] \geq \mathbb{E}\Big[\Big(q_i^{\mathsf{cs}}(\boldsymbol{X}|\boldsymbol{S}, f_i^{\mathsf{OR}}) - q_i^{\mathsf{Bayes}}(\boldsymbol{X})\Big)^2\Big],$$

while statement (b) implies that the above inequality holds for all i and for any fixed $\tau > 0$ and $\beta > 0$.

Proof of Lemma 5 - The proof of this Lemma follows directly using Theorem 1 for the numerator of $\hat{f}_i^{\mathsf{prop}}$ and Lemma 6 and equations (9), (10) for the denominator. Similar arguments using Theorem 3, Lemma 6 and equations (9), (10) prove the result for $\hat{f}_i^{\mathsf{prop}}$ in definition 4.

A.8 Proofs of Theorems 2 and 4

We will first prove Theorem 4 for the generalized absolute loss function in equation (3).

For any i and fixed $\tau > 0$, $\beta > 0$, we have ,

$$\Big|\hat{q}_i^{\mathsf{caspr}}(\hat{f}_i^{\mathsf{prop}}) - \hat{q}_i^{\mathsf{cs}}(f_i^{\mathsf{OR}})\Big| = \Big|\hat{f}_i^{\mathsf{prop}} - f_i^{\mathsf{OR}}\Big|\Big|\boldsymbol{e}_i^T \hat{H}_{1,-1,\beta}\boldsymbol{A}(\boldsymbol{X} - \boldsymbol{\eta}_0)\Big|,$$

which can be upper bounded by $\Big|1-\hat{f}_i^{\mathsf{prop}}/f_i^{\mathsf{OR}}\Big|\Big[\Big|\hat{q}_i^{\mathsf{cs}}(f_i^{\mathsf{OR}})-\boldsymbol{e}_i^T\boldsymbol{A}\boldsymbol{\eta}_0\Big|+\Big|\Phi^{-1}(\tilde{b}_i)\Big|\Big(\boldsymbol{e}_i^T\hat{G}_{1,0,\beta}\boldsymbol{e}_i+ m_0^{-1}\boldsymbol{e}_i^T\hat{G}_{0,1,0}\boldsymbol{e}_i\Big)^{1/2}\Big]$. Now using Theorem 3 and Lemma 5, $\|\hat{\boldsymbol{q}}^{\mathsf{caspr}}(\boldsymbol{A}\boldsymbol{X}|\hat{\boldsymbol{f}}^{\mathsf{prop}})-\hat{\boldsymbol{q}}^{\mathsf{cs}}(\boldsymbol{A}\boldsymbol{X}|\boldsymbol{f}^{\mathsf{OR}})\|_2^2$ is upper bounded by

$$\frac{2}{\Big(f_{\mathsf{inf}}^{\mathsf{OR}}\Big)^2}\Big[\|\hat{\boldsymbol{q}}^{\mathsf{cs}}(\boldsymbol{A}\boldsymbol{X}|\boldsymbol{f}^{\mathsf{OR}})-\boldsymbol{e}_i^T\boldsymbol{A}\boldsymbol{\eta}_0\|_2^2+\Big\{\Phi^{-1}(\tilde{b}_i)\Big\}^2\Big(\boldsymbol{e}_i^T G_{1,0,\beta}\boldsymbol{e}_i+m_0^{-1}\boldsymbol{e}_i^T G_{0,1,0}\boldsymbol{e}_i+c_n\Big)\Big]O_p\Big(\frac{\log n}{n}\Big),$$

where $f_{\mathsf{inf}}^{\mathsf{OR}} \coloneqq \inf_{1\le i\le n} f_i^{\mathsf{OR}} > 0$ and $c_n = O_p\Big\{\max\Big(\frac{p}{n}, \sqrt{\frac{\log n}{n}}\Big)\Big\}$. The proof then follows by noting that $\Big|\Big|\hat{\boldsymbol{q}}^{\mathsf{cs}}(\boldsymbol{A}\boldsymbol{X}|\boldsymbol{f}^{\mathsf{OR}})-\boldsymbol{e}_i^T\boldsymbol{A}\boldsymbol{\eta}_0\Big|\Big|_2^2 > 0$ since $\ell_0 > 0$. The proof of Theorem 2 follows using similar arguments as Theorem 1 and Lemma 5.

References

J. Aitchison and I. R. Dunsmore. Statistical prediction analysis. *Bulletin of the American Mathematical Society*, 82(5):683–688, 1976.

Jinho Baik and Jack W Silverstein. Eigenvalues of large sample covariance matrices of spiked population models. *Journal of Multivariate Analysis*, 97(6):1382–1408, 2006.

Gah-Yi Ban and Cynthia Rudin. The big data newsvendor: Practical insights from machine learning. *Operations Research*, 67(1):90–108, 2019.

Sudipto Banerjee, Bradley P Carlin, and Alan E Gelfand. *Hierarchical modeling and analysis for spatial data*. CRC press, 2014.

Zhigang Bao, Xiucai Ding, Jingming Wang, and Ke Wang. Statistical inference for principal components of spiked covariance matrix. *arXiv preprint arXiv:2008.11903*, 2020.

Florent Benaych-Georges and Raj Rao Nadakuditi. The singular values and vectors of low rank perturbations of large rectangular random matrices. *Journal of Multivariate Analysis*, 111:120–135, 2012.

Alex Bloemendal, Antti Knowles, Horng-Tzer Yau, and Jun Yin. On the principal components of sample covariance matrices. *Probability theory and related fields*, 164(1-2): 459–552, 2016.

Bart J Bronnenberg, Michael W Kruger, and Carl F Mela. Database paper—the iri marketing data set. *Marketing science*, 27(4):745–748, 2008.

Bart J Bronnenberg, Jean-Pierre H Dubé, and Matthew Gentzkow. The evolution of brand preferences: Evidence from consumer migration. *American Economic Review*, 102(6): 2472–2508, 2012.

Lawrence D Brown and Eitan Greenshtein. Nonparametric empirical bayes and compound decision approaches to estimation of a high-dimensional vector of normal means. *The Annals of Statistics*, pages 1685–1704, 2009.

Lawrence D Brown, Gourab Mukherjee, and Asaf Weinstein. Empirical bayes estimates for a 2-way cross-classified additive model. *Annals of Statistics*, 2018.

T Tony Cai, Zongming Ma, and Yihong Wu. Sparse pca: Optimal rates and adaptive estimation. *The Annals of Statistics*, 41(6):3074–3110, 2013.

Marielle Cavrois, Trambak Banerjee, Gourab Mukherjee, Nandhini Raman, Rajaa Hussien, Brandon Aguilar Rodriguez, Joshua Vasquez, Matthew H Spitzer, Nicole H Lazarus, Jennifer J Jones, et al. Mass cytometric analysis of hiv entry, replication, and remodeling in tissue cd4+ t cells. *Cell reports*, 20(4):984–998, 2017.

Olivier Coibion, Yuriy Gorodnichenko, and Gee Hee Hong. The cyclicality of sales, regular and effective prices: Business cycle and policy implications. *American Economic Review*, 105(3):993–1029, 2015.

Dipak K Dey, Sujit K Ghosh, and Bani K Mallick. *Generalized linear models: A Bayesian perspective*. CRC Press, 2000.

Lee H Dicker and Sihai D Zhao. High-dimensional classification via nonparametric empirical bayes and maximum likelihood inference. *Biometrika*, 103(1):21–34, 2016.

Bradley Efron. *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*, volume 1. Cambridge University Press, 2012.

Bradley Efron and Trevor Hastie. *Computer age statistical inference*, volume 5. Cambridge University Press, 2016.

Jianqing Fan, Yingying Fan, and Jinchi Lv. High dimensional covariance matrix estimation using a factor model. *Journal of Econometrics*, 147(1):186–197, 2008.

Jianqing Fan, Yuan Liao, and Martina Mincheva. Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(4):603–680, 2013.

Jianqing Fan, Weichen Wang, and Yiqiao Zhong. Robust covariance estimation for approximate factor models. *Journal of econometrics*, 208(1):5–22, 2019.

Dominique Fourdrinier, William E. Strawderman, and Martin T. Wells. *Shrinkage Estimation*. Springer, 2017.

S Geisser. Predictive inference. monographs on statistics & applied probability. *CRC, London*, 1993.

Andrew Gelman and Jennifer Hill. *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press, 2006.

Edward I George and Xinyi Xu. Predictive density estimation for multiple regression. *Econometric Theory*, 24(2):528–544, 2008.

Edward I George, Feng Liang, and Xinyi Xu. Improved minimax predictive densities under kullback–leibler loss. *The Annals of Statistics*, 34(1):78–91, 2006.

John Geweke. The dynamic factor analysis of economic time series. *Latent variables in socio-economic models*, 1977.

Linda V Green, Sergei Savin, and Nicos Savva. "nursevendor problem": Personnel staffing in the presence of endogenous absenteeism. *Management Science*, 59(10):2237–2256, 2013.

Eitan Greenshtein and Junyong Park. Application of non parametric empirical bayes estimation to high dimensional classification. *Journal of Machine Learning Research*, 10 (Jul):1687–1704, 2009.

Eitan Greenshtein and Ya'acov Ritov. Asymptotic efficiency of simple decisions for the compound decision problem. In *Optimality: The Third Erich L. Lehmann Symposium*, volume 57, pages 266–275, 2009.

Campbell R Harvey, Yan Liu, and Heqing Zhu. ... and the cross-section of expected returns. *The Review of Financial Studies*, 29(1):5–68, 2016.

Iain M Johnstone and Arthur Yu Lu. On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association*, 2012.

Iain M Johnstone and Debashis Paul. Pca in high dimensions: An orientation. *Proceedings of the IEEE*, 106(8):1277–1292, 2018.

Iain M Johnstone and D Michael Titterington. Statistical challenges of high-dimensional data, 2009.

Noureddine El Karoui and Holger Kösters. Geometric sensitivity of random matrix results: consequences for shrinkage estimators of covariance and related statistical methods. *arXiv preprint arXiv:1105.1404*, 2011.

Noureddine El Karoui, Andrew EB Lim, and Gah-Yi Vahn. Estimation error reduction in portfolio optimization with conditional value-at-risk. Technical report, 2011.

Koulik Khamaru and Rahul Mazumder. Computation of the maximum likelihood estimator in low-rank factor analysis. *Mathematical Programming*, 176(1):279–310, 2019.

Roger Koenker and Gilbert Bassett Jr. Regression quantiles. *Econometrica: journal of the Econometric Society*, pages 33–50, 1978.

Roger Koenker and Ivan Mizera. Convex optimization, shape constraints, compound decisions, and empirical bayes rules. *Journal of the American Statistical Association*, 109 (506):674–685, 2014.

Fumiyasu Komaki. Asymptotic properties of bayesian predictive densities when the distributions of data and target variables are different. *Bayesian Analysis*, 10(1):31–51, 2015.

SC Kou and Justin J Yang. *Optimal shrinkage estimation in heteroscedastic hierarchical linear models*. Springer, 2017.

Serhiy Kozak, Stefan Nagel, and Shrihari Santosh. Shrinking the cross-section. *Journal of Financial Economics*, 135(2):271–292, 2020.

Shira Kritchman and Boaz Nadler. Determining the number of components in a factor model from limited noisy data. *Chemometrics and Intelligent Laboratory Systems*, 94(1): 19–32, 2008.

Shira Kritchman and Boaz Nadler. Non-parametric detection of the number of signals: Hypothesis testing and random matrix theory. *IEEE Transactions on Signal Processing*, 57(10):3930–3941, 2009.

Zongming Ma. Sparse principal component analysis and iterative thresholding. *The Annals of Statistics*, 41(2):772–801, 2013.

Gourab Mukherjee, Lawrence D Brown, and Paat Rusmevichientong. Efficient empirical bayes prediction under check loss using asymptotic risk estimates. *arXiv preprint arXiv:1511.00028*, 2015.

Alexei Onatski. Asymptotics of the principal components estimator of large factor models with weakly influential factors. *Journal of Econometrics*, 168(2):244–258, 2012.

Art B Owen and Jingshu Wang. Bi-cross-validation for factor analysis. *Statistical Science*, 31(1):119–139, 2016.

Damien Passemier and Jian-Feng Yao. On determining the number of spikes in a high-dimensional spiked population model. *Random Matrices: Theory and Applications*, 1 (01):1150002, 2012.

Damien Passemier, Zhaoyuan Li, and Jianfeng Yao. On estimation of the noise variance in high dimensional probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2015.

Damien Passemier, Zhaoyuan Li, and Jianfeng Yao. On estimation of the noise variance in high dimensional probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(1):51–67, 2017.

L'uboš Pástor. Portfolio selection and asset pricing models. *The Journal of Finance*, 55(1): 179–223, 2000.

L'uboš Pástor and Robert F Stambaugh. Comparing asset pricing models: an investment perspective. *Journal of Financial Economics*, 56(3):335–381, 2000.

Andrew J Patton and Allan Timmermann. Properties of optimal forecasts under asymmetric loss and nonlinearity. *Journal of Econometrics*, 140(2):884–918, 2007.

Debashis Paul. Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistica Sinica*, pages 1617–1642, 2007.

Debashis Paul and Alexander Aue. Random matrix theory in statistics: A review. *Journal of Statistical Planning and Inference*, 150:1–29, 2014.

S James Press. *Subjective and objective Bayesian statistics: principles, models, and applications*, volume 590. John Wiley & Sons, 2009.

Herbert Robbins. The empirical bayes approach to statistical decision problems. In *Herbert Robbins Selected Papers*, pages 49–68. Springer, 1985.

Nandini Sen, Gourab Mukherjee, Adrish Sen, Sean C Bendall, Phillip Sung, Garry P Nolan, and Ann M Arvin. Single-cell mass cytometry analysis of human tonsil t cell remodeling by varicella zoster virus. *Cell reports*, 8(2):633–645, 2014.

Wenguang Sun and T Cai. Large-scale multiple testing under dependence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(2):393–424, 2009.

Zhiqiang Tan. Improved minimax estimation of a multivariate normal mean under heteroscedasticity. *Bernoulli*, 21(1):574–603, 2015.

Hal R Varian. A bayesian approach to real estate assessment. *Studies in Bayesian econometrics and statistics in honor of Leonard J. Savage*, pages 195–208, 1975.

Asaf Weinstein, Zhuang Ma, Lawrence D Brown, and Cun-Hui Zhang. Group-linear empirical bayes estimates for a heteroscedastic normal mean. *Journal of the American Statistical Association*, 113(522):698–710, 2018.

Xianchao Xie, SC Kou, and Lawrence D Brown. Sure estimates for a heteroscedastic hierarchical model. *Journal of the American Statistical Association*, 107(500):1465–1479, 2012.

Xianchao Xie, Samuel C Kou, and Lawrence Brown. Optimal shrinkage estimation of mean parameters in family of distributions with quadratic variance. *Annals of statistics*, 44(2): 564, 2016.

Keisuke Yano and Fumiyasu Komaki. Information criteria for prediction when the distributions of current and future observations differ. *Statistica Sinica*, pages 1205–1223, 2017.

Arnold Zellner. Bayesian estimation and prediction using asymmetric loss functions. *Journal of the American Statistical Association*, 81(394):446–451, 1986.

Cun-Hui Zhang. Compound decision theory and empirical bayes methods: invited paper. *Ann. Statist.*, 31(2):379–390, 04 2003.